**Queen Mary University of London**
School of Electronic Engineering
and Computer Science

# Human Expressivity in the Control and Integration of Computationally Generated Audio

**Christian Heinrichs**

**PhD Thesis**

Primary Supervisor: Dr Andrew McPherson
Centre for Digital Music
Queen Mary University of London

Second Supervisor: Dr Josh Reiss
Centre for Digital Music
Queen Mary University of London

Submitted in partial fulfilment of the requirements of the
University of London for the Degree of Doctor of Philosophy

**Abstract**

While physics-based synthesis offers a wide range of benefits in the real-time generation of sound for interactive environments, it is difficult to incorporate nuanced and complex behaviour that enhances the sound in a narrative or aesthetic context. The work presented in this thesis explores real-time human performance as a means of stylistically augmenting computational sound models. Transdisciplinary in nature, this thesis builds upon previous work in sound synthesis, film sound theory and physical sound interaction. Two levels on which human performance can enhance the aesthetic value of computational models are investigated: first, in the real-time manipulation of an idiosyncratic parameter space to generate unique sound effects, and second, in the performance of physical source models in synchrony with moving images. In the former, various mapping techniques were evaluated to control a model of a creaking door based on a proposed extension of practical synthesis techniques. In the latter, audio post-production professionals with extensive experience in performing Foley were asked to perform the soundtrack to a physics-based animation using bespoke physical interfaces and synthesis engines. The generated dataset was used to gain insights into stylistic features afforded by performed sound synchronisation, and potential ways of integrating them into an interactive environment such as a game engine.

Interacting with practical synthesis models that have extended to incorporate performability enables rapid generation of unique and expressive sound effects, while maintaining a believable source-sound relationship. Performatively authoring behaviours of sound models makes it possible to enhance the relationship between sound and image (both stylistically and perceptually) in ways precluded by one-to-one mappings between physics-based parameters. Mediation layers are required in order to facilitate performed behaviour: in the design of the model on one hand, and in the integration of such behaviours into interactive environments on the other. This thesis provides some examples of how such a system could be implemented. Furthermore, some interesting observations are made regarding the design of physical interfaces for performing environmental sound, and the creative exploitation of model constraints.

# Acknowledgements

I would like, first and foremost, to thank my supervisor Dr Andrew McPherson. I am forever grateful for the guidance, enthusiasm, support, patience and expertise he has lent me throughout this project, as a mentor, a teacher and a friend. My deep gratitude is extended to Dr Andy Farnell (my 'pub supervisor'), without whom this thesis would never have come to fruition. The countless discussions we have had - on, around and beyond the topic of computational audio - have been a huge inspiration. I would also like to thank Dr Josh Reiss for always offering interesting and alternative perspectives, particularly during the progression review meetings for this project.

Thanks to Richard Kelly, for the tremendous amount of work he put into making the Media and Arts Technology PhD programme what it is today; to Kok Ho Huen, Jonathan Winfield and Geetha Bommireddy for their technical support and for running the labs; Giulio Moro and Chris Harte for always going the extra mile in running the MAT studios.

Thanks are due to Graham Gatheral, for establishing the *Procedural Audio Now!* and Game Audio meetups, and for his input into some of the early prototypes for the second study; to John Roesch for going out of his way to provide deep insights into the art of Foley; to Robert 'Diz-B' Thomas, Joe White and Martin Roth, for all the deep discussions and practical context they have provided; to David Lees and Stefanie Ritch for creating the radio play for my first study (and achieving the seemingly impossible task of making a script revolving around squeaky doors entertaining); to all the participants of my studies - this work would not have been possible without you.

Thank you to everybody on the MAT programme for fostering an inspiring community of researchers, artists and friends; and to everybody in C4DM. Above all, it has been an honour to witness the Augmented Instruments Lab grow into the incredibly energetic and enthusiastic research group that is has become today - keep dialing it up to eleven!

My gratitude is extended to Robert Jack, for having shared this long and ever-meandering journey into sound, music and electricity - I look forward to the many future creations that lay ahead of us! Thanks to Flora Pitrolo, Jonny Rogerson, Anya

Keenan, Helen Murphy and David Lees, for good times; to Victor Zappi for all the adventures in London and beyond; to Duncan Menzies and Laurel Pardue (two of the finest fiddle players I have encountered); to James Bradley and Olsen Wolf for putting things in their place; to Alessia Milo for all her support and inspiration.

I would also like to deeply thank everybody who helped me in the very final stages of this project. Thank you Chris Cunningham for showing me the light - quite literally - and for giving me confidence in the fact that art and creativity can and must always prevail over technology. These thanks are extended to Eddie Jeffrys, Steph Thirion, Tom Blackburn and Matt Davis, who supported me while I struggled to get the last words on paper inbetween all the lines of code.

Such words on paper are not enough to thank my flatmate, best friend and life partner, Pauline Serrano, for all the love, patience and inspiration she has given me throughout the final stages of this process. A personal journey may have ended for me, but I am grateful to know that the adventure that lies ahead of us is larger than life itself. I will take this moment to also thank our beloved Buttercat, who really helped me get through the long nights and never-ending mornings that went into this document.

Last, but very certainly not least, my infinite gratitude goes to my parents, Lorraine and Georg Heinrichs, and to my brother, Mark. I wouldn't be where I am now without all their encouragement, love and support. This work is dedicated to them.

This work was funded by the Engineering and Physical Sciences Research Council (EPSRC) as part of the Doctoral Training Centre in Media and Arts Technology at Queen Mary University of London (ref: EP/G03723X/1).

# Contents

# List of Figures

# List of Tables

# Related Publications

Christian Heinrichs, Andrew McPherson, and Andy Farnell. Human performance of computational sound models for immersive environments. *The New Soundtrack*, 4(2): 139–155, September 2014

Christian Heinrichs and Andrew McPherson. Mapping and Interaction Strategies for Performing Environmental Sound. In *IEEE VR Workshop: Sonic Interaction in Virtual Environments (SIVE)*, Minneapolis, MN, 2014

Christian Heinrichs and Andrew McPherson. Performance-Led Design of Computationally Generated Audio for Interactive Applications. In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 697–700. ACM Press, 2016

# Chapter 1

# Introduction

## 1.1 Motivation

Currently, the world of interactive audio is moving towards computational generation at runtime. Higher bandwidths of interaction in new media such as virtual reality require the soundtrack to respond with much greater degrees of nuance. This poses a great challenge to conventional sound design workflows, which rely heavily on pre-recorded sound and fixed waveforms to generate the interactive soundtrack.

In current workflows, providing adequate auditory feedback to complex and continuous interactions requires a large set of waveforms to account for all combinations of varying dimensions (for example, velocity and orientation of a virtual object). In contrast, a computational model of the given object or interaction is capable of producing a realistic and responsive soundtrack by simulating the underlying physical processes in real-time. This eliminates the need to record a large set of fixed waveforms, thus providing a significant boost in efficiency. However, current approaches to producing and implementing such models fall outside the expertise of a typical sound designer, requiring a robust understanding of physical processes, their mathematical formulations and means of implementing them digitally in a real-time context. More pertinently, sound qualities that would normally be left to the sound designer to craft are no longer in their control but rather the result (or by-product) of a simulation. For example, a detailed physical model of stick-slip friction makes it possible to simulate the sound of squealing brakes and creaking hinges, but also makes it difficult to predict the resulting sonic behaviour (for example, when attempting to produce a particularly 'spooky' or 'squeaky' door sound).

At its core, sound design is a discipline that relies on experimentation and fast iteration, guided by a subjective listening process in order to provide a meaningful soundtrack for an extra-auditory context. Physical interaction is often central to this

process. For example, a pair of leather gloves might be used to dramatically emulate – *perform* – the sound of birds taking off.

This work focuses on ways of bringing the human element back into computationally generated soundtracks. Bringing together the disciplines of sonic interaction design, digital musical instruments, sound synthesis and sound design for the moving image, this thesis seeks ways of incorporating human performance of sound into the design and integration of computational soundtracks.

### 1.1.1 Interactive Audio

To this day, the use of interactive audio in the entertainment industry has stood in contrast to developments in the fields of musical interaction and sonically guided interaction. Implementations in entertainment contexts, such as games, have relied on very simple interactional principles based on the triggering of pre-recorded waveforms. While this limits the responsiveness of sound to user interactions, fixed waveforms allow stylistic and aesthetic qualities inherited from motion picture sound to be easily transferred.

In contrast, current research in the design of digital musical instruments (DMIs) and sonic interaction design (SID) focuses on highly nuanced interaction with sound, which is often facilitated by the use of synthesis models based on physical simulations. However, the two fields can be seen to be in the process of converging. With the increasing quality and use of freehand motion controllers in games and the complexity of graphical animations, there is a need for more nuanced sound feedback that simultaneously fulfils the aesthetic requirements of the game. On the other hand, design principles developed in the fields of SID and sonification are starting to become adopted in the industry, where stylized sound is a highly important factor (e.g. for branding or reaching a particular target group). Game audio is thus an integral field here. It operates on the quality requirements of conventional cinema and thereby represents a working compromise between established audio-visual languages and the requirements of interactive media.

Some confusion may arise in that his work explores interaction with sound as a means of enhancing a further end-user interactive audio experience. As such there are two types of audio interactions at hand here: the *performative* use of sound as part of a design task, and the resulting interactive sound model that is deployed within its required context. Some of the ideas explored in the fields of DMI design and SID are relevant to both cases, and care has been taken to distinguish between the two contexts throughout this thesis. This topic is explored further in Section 2.4.8.

### 1.1.2 Computationally Generated Audio

Central to the advancement of interactive audio in all of the above fields is the computational generation of sound. By simulating sound sources and their underlying physical behaviours it is possible to generate a diverse and perceptually meaningful range of sounds in response to a small set of continuous, physically viable parameters. In the design of digital musical instruments (DMIs) this means that familiar acoustical systems (e.g. saxophones or vibrating bars) can be repurposed and extended to afford novel and expressive musical interaction techniques. In the field of SID, it has been applied to guide physical interactions with everyday objects by repurposing sonic affordances of an unrelated sound source. Throughout the remainder of this thesis the term *computationally generated audio* (CGA) will be used to refer to the creation of meaningful non-musical sound with little or no use of recorded material.Rather than the wider scope of sound synthesis, this term pertains directly to its application in professional sound design practice in the entertainment industry and the arts. Just as computer-generated imagery (CGI) provides digitally synthesised elements to an otherwise photography-based medium in cinema, CGA can be thought of as its sonic counterpart – where sound synthesis coexists as part of an otherwise microphone-oriented craft.

### 1.1.3 Sound Design and Aesthetics

Since the birth of cinema better recording and reproduction technology, and more sophisticated means of authoring sounds have resulted in the ability to design soundtracks with a high level of acoustic fidelity. As noted by numerous filmmakers and scholars (see for example (Weis and Belton, 1985)), the most valued aspects of the soundtrack rest in its contextual function, resulting in a somewhat loose relationship between sound and image. Thus, cultural implications are as crucial as the technological apparatuses by which they are created and reproduced. While realism and sound quality play an important role, priority is given to fulfilling audience expectations and heightening the flow of the narrative through dynamic (or *behavioural*) properties of the sound, which may therefore diverge from the physical action that the sound accompanies.

While computational models of sound sources provide a high degree of interactivity, they fail to take into account these divergences in a way that is conducive to the level of artistic control required by sound designers (Farnell, 2011). Instead, CGA is often approached from a purely technical perspective, with incentives to recreate sound that is physically realistic, computationally efficient and scalable (Mullan, 2010). This often requires a deep understanding of physical processes and numerical methods for achieving a corresponding simulation, with audible results only evaluable after the implementation of the model (Bilbao, 2009b).

This stands in stark contrast to approaches taken throughout the history of

sound design, where human performance and fast iteration cycles are central. Examples range from Foley artists performatively synchronising sounds to moving images, playable contraptions built by James MacDonald to recreate wind, thunder, croaks and many other sound effects for early animated cartoons, to the use of vocalisations to sketch or enhance elements of the soundtrack. Indeed, the use of expressive non-speech vocalisation has been shown to be widely used in conversational and everyday interactions, particularly in playful interactions with objects among children. Performance makes it possible to describe and emphasise behavioural qualities of sound without the direct requirement to provide a physically accurate representation of the associated source.

## 1.2 Objectives and Research Questions

This thesis seeks to explore and develop new methods for extending CGA to incorporate expressive stylistic qualities made possible by human performance. Two aspects of CGA are the primary focus of this thesis: *performed behaviour* in the deployment of computational models and *asynchronicity* of sound in relation to a moving image.

### 1.2.1 Performed Behaviour in the Deployment of CGA

Only recently have transdisciplinary approaches started to be considered in CGA, spearheaded by the work of Andy Farnell who under the umbrella-term *Procedural Audio* refers to 'sound as process' and a design methodology that encompasses 'more than physical modelling' and synthesis (Farnell, 2011). His proposed methodology is centred around a diverse range of analytical techniques (acoustical, physical and perceptual) to develop physical simplifications of sound sources using a design-oriented approach. Central to this approach is the concept of *behavioural abstraction* which aids the parameterisation of sound models, where top-level *physically viable* parameters are transformed into time-varying parameters of a lower-level signal chain. Physical viability refers to the ability of such parameters to easily map to a physical representation of space, energy, movement, and so forth – as in the employment of physical sensors for DMIs (see Section 2.4.4) and physics engines for games (see Section 2.2.2).

Chapters 3 and 4 explore and evaluate ways of extending Farnell's approach in order for a human to perform behavioural sequences in real-time (instead of encoding them procedurally into the model), while retaining a perceptually meaningful definition of the corresponding source. A new sub-class of *timbre-led models* is proposed, where sounds are produced by dynamically varying a set of parameters that correspond to perceptual instead of physical features. Strategies for performing such a

model were implemented based on established techniques in DMI design and evaluated in a user study.

### 1.2.2 Asynchronicity of Sound in Relation to the Moving Image

While the first stage of research addresses the performance of discrete sequences of sound using computational models, the final stage focuses on stylistic issues surrounding the *integration* of CGA into interactive environments.

Physics-based sound models are typically designed to be controlled using a small set of parameters that correspond to physical movements and states, which in turn can be extracted from corresponding simulations in an interactive environment. The result is a tightly coupled relationship between sound and image, surpassing levels of responsiveness achievable with sample-based approaches currently adopted in the games industry. This conflicts with stylistic principles found in cinema and other conventional media, where direct sound from the film-set is often rejected in favour of human interpretations of audible actions by a Foley artist. In fact, the stylistic use of ambiguity in the relationship between sound and moving images (or *asynchronicity*) has been considered to be a fundamental tool by film sound theorists and filmmakers ever since the introduction of sound to film (Eisenstein, 1987; Pudovkin, 1985).

A complex experimental environment has been developed to study stylistic strategies applied in the performed synchronisation of computational sound to moving images by Foley artists (presented in Chapter 5). The development of a physics-based graphical animation made it possible to analyse performed soundtracks with reference to objective descriptions extracted from the image. This included physical movement (that could directly drive the sound models in a conventional implementation) and higher-level events and states corresponding to the animation system. Observations presented in Chapter 6 can inform novel integration strategies for CGA that take performed stylistic deviations from the image into account.

### 1.2.3 Research Questions

Corresponding to the two areas of interest described above, this thesis seeks to answer the following research questions:

1. How can principles from practical sound synthesis be leveraged to enable the performative design (or real-time composition) of complex sound behaviours?

2. How do expressive sound synchronisation aesthetics and techniques for the moving image challenge current strategies for integrating computational audio into interactive media?

## 1.3 Thesis Structure

**Chapter 2** provides an overview of related work and the surrounding fields. Section 2.1 explores some key historical developments in the re-enactment of non-musical sound before the widespread availability of digital computers. Section 2.2 serves as an overview of practical synthesis in light of current techniques and workflows for deploying sound in games and other interactive audio applications. Section 2.3 introduces concepts and terms relevant to sound design aesthetics, encompassing film sound theory and electro-acoustic music. Section 2.4 focuses on real-time human interaction with computational audio, with a particular emphasis on the field of digital musical instruments.

**Chapter 3** presents a novel class of computational sound models named *timbre-led models*. As this is an extension to the *practical synthesis* design strategy proposed by Farnell (2008), his approach is summarised in Section 3.1. Particular focus is placed on what Farnell refers to as the *behavioural* qualities of the sound, and how they are represented in the architecture of the sound model. Section 3.2 presents a new proposed class of *timbre-led models*, in which the behavioural parameter space of the sound model is replaced with a perceptual one.

**Chapter 4** is centred around the introduction of human performance to rapidly design behaviours for a timbre-led model of a creaking door. Section 4.2 provides a brief overview of stick-slip friction (the physical process underlying the sound associated with creaking doors) and ways of modelling it on both a source and signal basis. Section 4.3 describes the development process of a timbre-led creaking door model, starting with a basic signal model developed by Farnell (2008) and extending it to incorporate increasingly complex sonic features observed in recorded reference material. Following the design strategy outlined in Chapter 3, each iteration of the development results in an additional parameter (or set of parameters) that can be used to vary the sound output in real-time. Section 4.4 presents three different control layers that transform sensor data from a touch-capacitive surface to three varying parameters of the creaking door model. These include *one-to-one* and *many-to-many* mappings (as commonly applied in HCI and in digital musical instruments) and a novel mapping strategy referred to here as a *physically-inspired control layer*, which emulates the physical behaviour of a bowed string. In addition to this, four metrics for evaluating the performable model (including the aforementioned control layers) are presented: *range*, *nuance*, *repeatability* and *believability*. Finally, Section 4.5 presents the procedure and results of a user study evaluating the performable model involving 15 participants with experience in sound design and musical performance.

**Chapter 5** shifts the focus to the synchronisation of sound to continuous visual movement. Section 5.1 provides a brief overview of current integration strategies for generating real-time soundtracks for interactive environments, including *one-to-one* mappings of physical data to sound models, the event-sample paradigm that is currently widely adopted in the games industry and potential approaches for computational models based on parameter sequences proposed in the previous chapters. Section 5.2 presents results from a survey conducted with professional Foley artists regarding aspects of performance and synchronisation that they deem important in a conventional motion picture context. Section 5.3 describes an experimental environment that was developed to study sound-image relationships as performed by professional Foley practitioners using physical interfaces driving computational models.

**Chapter 6** presents the procedure and results of a study based on the experimental environment described in Section 5.3. Six participants with extensive experience in performing Foley were asked to develop a soundtrack for a two-minute animation using a set of sound models and corresponding physical interfaces. The resulting soundtracks were later compared to a reference soundtrack generated by physical reference data extracted from the animation. Section 6.1 provides a technical overview of the configuration and procedure of the study. Section 6.2 summarises general observations made in regard to the animation, sound models and physical interfaces featured in the study. Section 6.3 provides an overview of how soundtracks were structured by the participants, with a particular focus on their categorisation of sounds across separately performed *takes* or *layers*. Section 6.4 explores some of the emerging patterns and themes observed across participants' soundtracks. Section 6.5 proposes potential integration strategies based on five brief case studies from the performed soundtracks. Finally, some conclusions are drawn in Section 6.6, which are discussed in more detail in Chapter 7.

**Chapter 7** provides a broader discussion of the findings from the previous chapters, considering both their technical and aesthetic implications in the design and integration of CGA for interactive environments. Section 7.1 explores some of the technical considerations involved in integrating discrete sequences of performed behaviours as featured in Chapters 3 and 4. Parallels are drawn to graphical animation, where motion of virtual object representations is typically expressed as separate interpolatable sequences of data. Section 7.2 reviews the key findings of the final synchronisation study, highlighting ways in which basic interpolation techniques fall short, rendering both the event-sound paradigm and the one-to-one mapping of movement to sound inadequate for the meaningful integration of performed soundtracks. Finally, Section

7.3 opens the discussion to broader issues including unexpected findings that warrant further investigation in future work.

**Chapter 8** concludes the thesis by summarising the key contributions made and presenting an overview of future directions of research.

# Chapter 2

# Background

Transdisciplinary in nature, this research project ties together several different fields within three core disciplines of sound design, sound synthesis and interaction with computational audio. An exhaustive overview of each of these fields falls beyond the scope of this thesis. Instead some fundamental concepts and themes that overlap across disciplines and are relevant to this work will be presented here.

Section 2.1 explores some key historical developments in the re-enactment of non-musical sound before the widespread availability of digital computers. Section 2.2 provides an overview of computational sound synthesis techniques including physical modelling and practical approaches that take computational or perceptual limitations into account. Their benefits and shortcomings are discussed further in the light of current approaches to game audio. Section 2.3 introduces concepts and terms relevant to sound design aesthetics, encompassing film sound theory and electro-acoustic music. Section 2.4 focuses on real-time human interaction with computational audio, with a particular emphasis on the field of digital musical instruments.

## 2.1 Imitation of Non-Musical Sound before Computers

Most people have reenacted sounds from their everyday environment, for example the sounds of a loud train rushing past, earth-shattering thunder or a mosquito buzzing past one's ear. Onomatopoeia have long been used and quickly grasped in literature, comic books and children's stories, such as the 'KAPOW!', 'WHAMM!' and 'SPLOOSH' inserts in the 1960s television series *Batman*, or verses in many of the popular poems by Dr Seuss. Humans have a natural desire to exaggerate, parody or otherwise embellish sound to disclose a particular effect or attribute of the referenced sounding object.

Though this work deals primarily with digital representations of sound it is worth giving a quick overview of what one could consider as the beginnings of synthetic (i.e. imitated or emulated) environmental sound before the rise of pervasive computing. While the human voice is commonly used to great effect to imitate sounds in everyday situations, some have built contraptions to recreate environmental sounds and Foley artists continue to this day to imitate sounds for moving images using both similar or unrelated sound-producing objects.

### 2.1.1 Onomatopoeia and Vocal Imitation

Onomatopoeia are words or utterances that imitate a sound to refer to the object that has caused it. In his seminal work 'Prehistoric Man' Wilson (1865) illustrates the different vocalisations of the sound of a drum in different languages: dundubhi (Sanscrit), rubadub (Greek), rattaplan (French), trumberum (German), czimbalom (Hungarian), tung-tung (Manchu), kan-kan (Chinese). He later describes different accounts of the cry of a whippoorwill (Caprimulgus vociferus): kwa-kor-yeuh, wha-oo-nah, eh-poo-weh and of course as it is named in English - whip-poor-will.

Some onomatopoeia allow for various degrees of expression to be encoded in the transmitted sound. For example, one might consider the different ways of vocalising an explosion. The common utterance - 'boom!' - can be expressed at different volume levels, the downward inflection of the vowel can be made more extreme by starting at a higher pitch, the ending consonant can be replaced with 'shhh' to simulate a long reverberant tail, or a prefix can be added to emphasise the sound's attack (e.g. '*ka*-boom!'). While the original version would suffice for the simple purpose of referral, these and various other forms of embellishment can be used to communicate other attributes such as the severity of the explosion and its evolution over time.

Lemaitre and Rocchesso (2014) distinguish between onomatopoeia and *non-conventional or creative* vocal imitations. While onomatopoeia are integrated into a language as words (therefore having a direct semantic referral to the object), direct vocal imitations of sources can also be used to refer to objects and associated actions. The use of creative vocalisations to accompany playful interactions with the environment has been observed in children from a very young age (Dumaurier et al., 1982). Vocal imitations illustrate the ability to convey properties of a sounding object using a limited and mechanically unrelated sound synthesis apparatus. Most importantly, sound is re-enacted in this way not only as a means of referral but as a way of conveying extra layers of meaning associated with the sound or one's relationship to the sounding object. A more detailed overview of the subject is provided in Lemaitre and Rocchesso (2014).

### 2.1.2 Russolo and Futurism

At the turn of the 20th century the explosion of industrialism gave rise to a new art movement in Italy known as Futurism. Painters and sculptors instigated the movement, portraying their vision of a new world inhabited by oppressive machines and rapid sterile movements. Inspired by the industrial noises that started to invade the everyday soundscape, the infamous futurist painter Luigi Russolo envisioned a new form of music based on these sounds. He criticised modern contemporary music for its dependence on consonant sounds and conventional forms. In a detailed manifesto addressed to the composer Balilla Pratella he proposed to expand the sonic palette of music to what he perceived as the fascinating noises made by the machines of the new industrial age (Russolo, 1913):

> Let us cross a great modern capital with our ears more alert than our eyes, and we will get enjoyment from distinguishing the eddying of water, air and gas in metal pipes, the grumbling of noises that breathe and pulse with indisputable animality, the palpitation of valves, the coming and going of pistons, the howl of mechanical saws, the jolting of a tram on its rails, the cracking of whips, the flapping of curtains and flags. We enjoy creating mental orchestrations of the crashing down of metal shop blinds, slamming doors, the hubbub and shuffling of crowds, the variety of din, from stations, railways, iron foundries, spinning wheels, printing works, electric power stations and underground railways.

He concluded his manifesto, entitled 'The Art of Noises', by outlining six 'families' of noises that would account for every imaginable sound. They 'encapsulated the most characteristic of the fundamental noises; the others are merely the associations and combinations of these' (Russolo, 1913) (see Table 2.1). It is worth noting that, in most cases, the language used here refers to perceptual characteristics of *actions*, rather than the physical sources that originally generated the sounds. This stands in contrast to modern physics-based synthesis techniques (described below in Section 2.2), which typically focus on simulating virtual representations of sound sources rather than their associated behaviours.

In 1913 he invented and constructed a new family of instruments called Intonarumori to accommodate his musical ambitions. Each instrument consisted of a hollow wooden box containing a string attached to a drumhead mounted, which in turn was mounted onto a metal speaker. Turning a crank would cause the string to vibrate while a lever at the top of the instrument varied the tension of the string. The Intonarumori were built in different sizes and variations accommodating all members of his six 'families' of environmental sound outlined at the end of his manifesto (Russolo, 1913).

| | |
|---|---|
| 1 | Rumbles, Roars, Explosions, Crashes, Splashes, Booms |
| 2 | Whistles, Hisses, Snorts |
| 3 | Whispers, Murmurs, Mumbles, Grumbles, Gurgles |
| 4 | Screeches, Creaks, Rumbles, Buzzes, Crackles, Scrapes |
| 5 | Noises obtained by percussion on metal, wood, skin, stone, terracotta, etc. |
| 6 | Voices of animals and men: Shouts, Screams, Groans, Shrieks, Howls, Laughs, Wheezes, Sobs |

Table 2.1: Russolo's proposed *six families of noises* constituting the design of his Intonarumori (taken from Russolo (1913)).

Russolo continued to be an inspiration to new and aspiring composers such as Edgard Varèse and Iannis Xenakis, together forming an influential sub-genre of art music inspired by the structures underlying everyday sounds. Serafin et al. (2006) designed a physical controller and synthesis environment that emulated the interactive and sounding properties of the Intonarumori family (revisited in Section 2.4.10).

### 2.1.3 James MacDonald's Sound Effect Machines and Foley Artistry

A few decades later, while cinema was still regarded as a purely visual medium, it became possible to add a fixed soundtrack to film, giving rise to the 'talkies' (Weis and Belton, 1985). While in traditional films it was possible to conceal microphones on the film set to obtain all the key elements of the soundtrack, animated films had to rely on specialised composers and orchestras to generate the entire soundtrack with no reference sound material. Because orchestral scores played such a big role, percussionists were initially given the task of producing the majority of non-musical sound effects such as crashing plates, thunder, creaking doors, and so forth. One of these percussionists was James MacDonald who worked on some of Disney's early films. MacDonald presented increasingly innovative solutions for humorously re-enacting the comical actions on the screen, acquiring the attention of the animation studio's executives (Ament, 2009). MacDonald became known for his innovative contraptions that could recreate all manner of environmental sounds.

One of the first such machines MacDonald created consisted of a large hollow cylinder with finishing nails protruding along the inner surface. The cylinder was filled with dry peas and when rotated would produce the sound of falling rain. A variation of this device was capable of producing other water-based sounds from light streams to crashing waves. Other instruments included a stick attached to a tin can, which when bowed would produce the sound of a creaking frog, and huge sheets of metal which were used to perform the sound of rumbling thunder.

Foley art describes the performative use of physical objects (including shoes and props from a film set) to generate a soundtrack to a sequence of moving images. It is named after Jack Foley who was the first to perform sound effects in synchrony to a film due to limitations of sound recording technology. An insightful account of the history of Foley art can be found in (Ament, 2009), which is the only comprehensive overview of what is otherwise commonly referred to as a 'dark art' practised behind the closed doors of audio post-production studios. Nuances of Foley art practice are explored in more detail in Chapter 5, which includes results and testimonies from a survey conducted with 26 practitioners with significant experience in performing Foley.

MacDonald and many other early innovators - including our prehistoric ancestors - demonstrate, maybe through lack of other suitable resources, the ability to synthesise compelling sounds using unrelated objects. These objects are designed with two purposes in mind: synthesis and performance. Each of these devices possess the ability to transform some form of human movement into sound. They existed because of an interest in disclosing meaning through the sound; specifically, meaning that is easier to disclose through everyday actions such as vocalising, spinning, grinding, and so on. In a similar vein to onomatopoeic utterances, the objective here is not always just to emulate a given sound (for example to make a reference to its corresponding source), but also to render expressive human qualities, such as embarrassment in a person's gate or excessively drawn-out crashing sounds in slapstick films.

The next section shifts to modern, digital approaches that have been taken to imitate of sounds from the everyday environment.

## 2.2 Computationally Generated Audio

Over the last three decades two closely related approaches to the synthesis of environmental sound in the digital audio domain have gained traction: *physical modelling* and *practical synthesis*. While the former relies on mathematical representations of physical processes to model sound sources, the latter involves some degree of perceptual and/or artistic simplification while preserving physical constraints and behaviours. Practical synthesis has gained considerable attention in the realms of game audio and - to a lesser extent - in sound design for the stage. This is often referred to as *procedural audio*, coined by Farnell (2008) to underline the process-oriented approach to designing sound in contrast to the use of fixed samples. The merits of practical synthesis are usually recognized in its low memory requirements (in relation to sample-based playback) and relatively low computational cost (in relation to physical modelling) (Mullan, 2010). Less attention has been focused on stylisation and creative intervention in the way sound models are designed and integrated into virtual

environments or objects.

All of the above techniques will be referred to from hereon under the broader term of *computationally generated audio*. The next two sections provide a brief overview of physical modelling and practical synthesis, followed by a contextualisation of the two in light of contemporary sound design techniques and requirements.

### 2.2.1 Physical Modelling

It is possible to digitally approximate a wide variety of interactions with sounding objects by simulating the way that energy is dissipated through air and solid materials. This approach to synthesis is commonly referred to as *physical modelling*. Some of the primary applications of physical modelling are in virtual musical instruments, room acoustics and speech synthesis. Each method takes a different approach to the discretisation of time and space in the simulation of physical processes.

#### Lumped Models

*Lumped models* use virtual representations of mechanical units such as masses, springs and dampers to model the dissipation of energy across a spatial configuration. For example, by arranging multiple of these units in a linear spatial structure it is possible to simulate a string, and by extending this into a further dimension one can model a membrane. Lumped network models formed the basis of the CORDIS-ANIMA system developed by Cadoz et al. (1993) and implementations have recently been extended to account for haptic feedback and virtual reality interaction (Leonard et al., 2013).

#### Digital Waveguide Synthesis

Another technique is known as *digital waveguide* (DWG) synthesis, which approximates the propagation of waves through uniform tubes and strings through the use of delay lines. DWG synthesis emerged out of an abstract synthesis algorithm developed by Karplus and Strong (1983) producing timbres resembling a plucked string. Though originally formulated using principles of wavetable synthesis (Roads, 1996), the algorithm can be expressed as white noise passed through a feedback delay-line and a first-order averaging (low-pass) filter. Soon after, this synthesis technique was formalised by Jaffe and Smith (1983) as an efficient way of solving the wave equation in one dimension. While plucked-string applications of this method were developed further, particularly by Karjalainen et al. (1998), DWG synthesis has become a foundational technique in physical modelling and can be extended further to two or three dimensions to simulate membranes and room acoustics (Murphy et al., 2001). Complex spatial structures can be constructed out of numerous waveguides and scattering junctions, where the latter account for non-linear transformations at terminating

points of the corresponding waveguides.

### Modal Synthesis

In another commonly applied method known as *modal synthesis*, the simulation of sound sources is approached in the frequency domain. Vibrating objects are expressed in terms of their resonating modes, which are simulated using a bank of resonant filters or sinusoidal oscillators. Resonant modes are distinguished from spectrally complex onset sounds, which are typically used as an excitation source for the bank of resonators. While this technique is fairly straightforward in its signal implementation, simulating complex interactions requires an elaborate stage of analysis in order to deduce the modal parameters according to spatially varying properties (such as excitation position) (Van den Doel and Pai, 1996). A *shape-matrix* accounts for the variance of modal parameters (i.e. initial amplitudes, frequencies and damping coefficients) at different spatial points of the geometry. Ren et al. (2013) propose a perceptually-informed technique for deducing generic modal parameters of a given material by combining recordings of impacts with a virtual representation of the corresponding object. The method allows these parameters to later be transfered to new geometries.

### Numerical Solutions

More direct numerical simulations take their roots in applied mathematics and structural engineering. The behaviour of a sound source is modelled with partial differential equations (PDEs), which are solved at discrete points in space and time. Models can also account for changes in behaviour over time and are commonly referred to as *finite element time-domain* (FDTD) methods. The first known application of finite difference methods for purposes of sound synthesis was a simulation of vibrating strings by Hiller and Ruiz (1971). Since then, finite difference techniques have been applied in the simulation of many other sound producing components including reeds (Bilbao, 2009a), vibrating bars (Chaigne and Doutaut, 1997) and plates (Schedin et al., 1999).

Other numerical simulation methods include the finite element method and wave digital filters, which take their roots in structural analysis and the emulation of electrical circuits, respectively. A thorough overview of numerical approaches to sound synthesis can be found in (Bilbao, 2009b).

While finite difference simulations are typically computationally intensive in comparison to other methods, they are also highly parallelisable (in other words, the same PDE is solved across discrete points in space). A lot of current research focuses on exploiting GPU (graphical processor unit) architectures to accelerate computation times (for example, the recently concluded four-year *Next Generation Sound Synthesis* (NESS) project (Bilbao et al., 2014, 2013)). Some implementations are capable

of running in real-time at much lower spatial resolutions, such as the interactive two-dimensional simulation of wind instruments by Allen and Raghuvanshi (2015).

#### Audio-Visual Implementations

One of the focuses of physical modelling research is on the integration of physics-based sound synthesis into virtual graphical and interactive environments. Many such environments are based on integrated physics simulation engines. As these engines are centred around graphical representation, they do not support refresh rates required for numerical simulations in the audible or haptic range, or for physical behaviours associated with sounds that are perceptually pertinent but visually negligible. Therefore, audio-visual implementations either require approximating movement that is not accounted for in the existing simulation, or involve developing new physical simulations to drive both graphical and audible channels. Van den Doel and Pai (1996) attached a 'sound map' with corresponding modal parameters in order to simulate sounds in response to interactions (impacts) with virtual objects. Moss et al. (2010) use existing graphical fluid simulations to approximate the sound of water based on the amount and sizes of bubbles in each frame and an underlying sound model of a bubble that takes spherical harmonics into account.

More intricate models can be used to produce both sound and visual movement simultaneously, but require running physical simulations at much higher sampling rates, thus making them too computationally intensive for real-time implementations. O'Brien et al. (2001) simulated deformable bodies at high sampling rates using a finite element method to generate audio-visual images. Zheng and James (2011) present a similar simulation that takes more subtle effects into account, such as micro-collisions and chattering (e.g. plates audibly vibrating after a collision on its resting surface).

Chapters 5 and 6 focus on intentional divergences between visual movements and their audible counterparts in a narrative context, which would often be difficult or impossible to implement in the above approaches.

### 2.2.2 Practical Synthesis

The physical modelling techniques described above are capable of producing increasingly realistic results but usually require a lot of computational resources. In many cases this precludes the possibility for real-time implementations (using current commonly available processors). With incentives to incorporate physical simulations into interactive applications such as games and musical instruments, another strand of research has focused on developing strategies for simplifying simulations in order for them to run in real-time environments. These are referred to here under the broader term *practical synthesis*, in reference to a pragmatic approach that takes fac-

tors outside of pure acoustic simulation into account, such as computational efficiency, perceptual qualities or creative serendipity in the design process.

### Physically Informed Modelling

Cook (1997) introduces the concept of *physically informed* modelling, applying proposed techniques of *Physically Informed Spectral Additive Modelling* (PhISAM) and *Physically Informed Stochastic Event Modelling* (PhISEM) for percussive sounds. 'Physically informed' synthesis is characterised by a more lenient analysis process, allowing spectral and human analysis, with the aim of using the 'simplest synthesis models possible' while preserving physical consistency between input parameters and the resulting sound.

A subtle difference sets PhISAM apart from conventional modal synthesis techniques described above, which rely on physical analysis to deduce the parameters for a bank of resonant filters. Cook proposes a more pragmatic approach that can be led by combinations of spectral and human analysis, while retaining a physical parameterisation of the model. An advantage to the signal-oriented analysis is the ability to rapidly deduce new model configurations based on recordings without the need for a precise definition of the source's spatial and material properties. Excitation signals are produced using separate sub-models that take physical properties such as *stick hardness* and human interactional properties (e.g. *strike vigor*) into account.

PhISEM can be seen as an extension of granular synthesis that is driven by a simulation of physical particle systems. Velocities and collisions of point masses are calculated in order to derive parameters for sound synthesis of maracas, tambourines and police whistles. The simulation can be computed offline, deriving statistical distributions to drive granular synthesis parameters in real-time in response to control parameters such as *shake energy*.

Physically informed modelling strategies have been developed in both research and practical contexts, such as the water simulation model by Van den Doel (2005) and a wide range of models developed by Farnell (2008). Developments have also included the development of new real-time techniques based on physical modelling principles, such as *banded waveguides* presented by Essl et al. (2004). Hybrid methods that combine a range of physical and physically informed modelling techniques are applied in the Sound Design Toolkit (Delle Monache et al., 2010), to facilitate the exploration of novel interaction strategies in the field of Sonic Interaction Design. Models include a stick-slip friction model by Serafin (2004), which will be discussed in contrast to a novel approach presented in Chapter 4.

### Signal Models and the Unit Generator Paradigm

A design characteristic of physically informed models is the abstraction of the model into smaller sound or parameter generating components, allowing each component to be modelled separately. This process is highly compatible with the *unit generator* paradigm proposed by Mathews and Miller (1974). Unit generators refer to atomic signal processing blocks such as oscillators, filters and mathematical operations. They form the basis of many popular DSP authoring environments such as *Puredata*[1], *Supercollider*[2] and *Csound*[3]. Objectives behind the development of such environments include low latency (using small audio buffers), extensibility and rapid prototyping using graphical interfaces and the ability to modify processing graphs without interrupting the audio output (Van den Doel and Pai, 2001).

As each unit generator is a self contained component, signal processing *graphs* can be generated using both textual or graphical notations (e.g. blocks connected by virtual 'wires'). Processing graphs based on unit generators primarily express operations in the signal domain, typically operating on buffers of multiple samples in order to increase computational efficiency. The development of unit generator models is very different from numerical approaches, where translation into the signal domain usually occurs at the last stage of the simulation process.

### Applications in Virtual Environments

One of the key applications of practical synthesis methods is in the development of models for interactive virtual environments.

Van Den Doel et al. (2001) present methods of deducing parameters for impact and scraping models from a lower-rate physics engine (typically sampled at screen refresh rates between 25Hz and 120Hz). This stands in contrast to the implementation in a coinciding publication by O'Brien et al. (2001), where both visual images and sound are generated by running a physics simulation at a much higher rate (at or above typical audio sampling rates). Due to the simplicity of the models presented, multiple instances can be interacted with in real-time, and are easily integrated into existing commercial game engines. In contrast, the numerical approach by O'Brien et al. (2001) required between ninety and more than a thousand minutes to calculate a second of audio. These numbers would be much lower using current processors and optimisation methods, but the simulations would still be unlikely to run in real-time.

A more holistic framework was proposed in an influential paper by Takala and Hahn (1992). Under the term *sound rendering* a general framework for synchronising sounds to animations in virtual environments is proposed. A wide variety of sound

[1] http://puredata.info
[2] http://supercollider.github.io
[3] http://csound.github.io

reproduction methods are considered, ranging from real-time sampling of recorded sounds to physics-based and more abstract models. The main priority in each case is to achieve a close correlation to the visual animations by using movement and other extracted parameters to trigger events and drive parameters of synthesis models. Acoustic propagation in the virtually represented space is also considered. Many of the principles presented in this framework have become fundamental elements of modern interactive soundtracks for video games, for example the triggering of sound recordings by animations, the attachment of such sound generators to virtual objects and their propagation through virtual space (Collins, 2008). The use of real-time synthesis has been less prominent in video games, for reasons described below, but is steadily receiving more attention in the industry with the rise of consumer-grade virtual reality hardware and large virtual game universes.

A thorough framework for the implementation of real-time synthesis into virtual environments is provided by Farnell (2008), which has become more widely known under the term *procedural audio*. Concepts of abstraction and encapsulation using atomic components (introduced briefly above in the context of physically informed modelling) are taken a step further in an approach that is strongly influenced by object-oriented programming principles. Rather than primarily focusing on primitive objects such as plates, bars and tubes, complex mechanisms such as car engines, clocks and weather systems can be represented virtually by concatenating several lower-level models. Furthermore, such representations can be exploited to facilitate varying degrees of *level of audio detail* in order to improve computational efficiency in situations where multiple sound models and visual objects need to co-exist. For example, various parts of an aeroplane model could be disabled (or exchanged for computationally less demanding implementations) as it flies away from an observer into the distance. While this provides a useful solution to the computational cost of real-time synthesis, it also raises many opportunities for creative intervention. For example, the aeroplane model could blend into an existing musical background as it flies into the distance. Thus, computational audio can be approached with a consideration for aesthetic principles in parallel to physical and computational factors (Farnell, 2011).

#### Realism and Aesthetics

Realism and efficiency have been driving factors in the design of physical and practical models described above. However, to date there has been very little focus on aesthetic principles in the design and implementation of such models. As will be described in Section 2.3, introducing divergences between the soundtrack and the visual image is considered to be a crucial strategy for imparting meaning and aesthetic qualities in audio-visual media. Excessive focus on maintaining realism potentially limits the

scope in which computational models can be implemented (Farnell, 2014b).

A long overview of procedural audio by Mullan (2010) deals almost exclusively with the practicalities and technical milestones of the field. The main advantages are seen to be the low memory requirements and the infinite variations (in the sense of continuous multi-dimensional parameter spaces) of computationally generated sounds. In addition to the physics-based audio-visual systems described above, Menzies' *Phya* engine is a good benchmark representing this school of thought (Menzies, 2008), in which interactive objects are represented by sounds and graphics produced by the same physical algorithms. A more holistic, albeit technically less complex, implementation of synchronised audio-visual processes is presented by Verron in (Verron and Drettakis, 2012), where particle-based effects such as fire, wind and foliage are generated by the same forces within the game engine.

Implementations of CGA in the games industry (e.g. Wwise's SoundSeed[4], and AudioGaming's extensions of models described in (Farnell, 2008)[5]) tend to be limited to weather and particle-based effects, which easily replace their sample-based counterparts due to their sonic simplicity (ability to achieve recording-quality audio with simple signal chains) and ease of control (reactive and non-repetitive sounds with few controls and almost no memory footprint). Despite the common perception that procedural audio is capable of generating versatile audio that 'never sounds the same twice', a computational model can have the same issues of repetition as sample-based implementations if it is continually driven by the same animation or interaction (Farnell, 2011). When confronted on their scepticism towards procedural audio during a panel discussion at the *AES 49th Conference on Audio for Games* professional practitioners attributed it to opaque control interfaces and poor or sterile sound quality.

### 2.2.3 Game Audio and Sampling Techniques

While run-time sound rendering techniques such as acoustic propagation and event-based triggering of sounds have become increasingly sophisticated in the development of audio for games, the sonification of virtual sources is still largely based on techniques borrowed from motion picture sound.

Established techniques based on the recording and treatment of sound remain effective and irreplaceable tools for fixed film and animation media. As the soundtrack is usually produced and synchronised to the moving image at a later stage in production (*audio post-production*), sound designers can carefully assemble the soundtrack using hundreds of recordings (or *samples*) to match various properties of the visual image. These properties range from the psychological (how the viewer is supposed to feel) to the musical (how the sound blends in with the background music and/or

[4]http://www.audiokinetic.com
[5]http://www.audiogaming.net

ambience) to the acoustical (the spectral balance of the audio) and the physical (the believable representation of physical objects) (Farnell, 2014a; Collins, 2008).

While the interactive soundtracks of video games were originally synthesised in real-time with the aid of dedicated hardware components, the rapid increase of memory in portable storage formats resulted in the abandonment of synthesis in favour of sampled sound in the early 1990s. Recorded assets are programmed to be triggered and processed in response to events in the game, in a process commonly referred to as *integration* in the games industry (Collins, 2008).

While the sound quality of the samples themselves can easily match what is heard in cinema, the interactive nature of games leads to issues of repetition and unintended misalignment to visual movement and player interaction. Currently these issues are mitigated through the use of large sound libraries and the coarse manipulation of these samples by means of cross-fading, randomisation and granular sampling techniques. Audio engines (known as *audio middleware*) such as Wwise[6] and FMOD[7] are commonly used to handle these operations at runtime, while exposing interfaces for authoring the way in which they are triggered and processed in response to events and continuous parameters.

Advanced sampling techniques result in a slightly more varied sound image but are not enough to constitute a truly responsive environment. Farnell (2011) likens the resulting interactive soundtrack to the visual equivalent of a dynamic montage of photographs, where individual fragments exhibit high levels of 'surface realism' but don't correspond to a realistic or satisfying level of responsiveness (see Chapter 3 for a more detailed discussion). Due to the temporal linearity of film (and other fixed audio-visual media), sampled sound can be carefully composed to match the depicted movement, or to intentionally contradict it as an aesthetic choice. In non-linear media, however, the artifice of pre-recorded sound easily seeps into the consciousness of the spectator (or player). This can be mitigated somewhat by incorporating more audio content, however, the ever-increasing size of game universes and emergence of high-bandwidth control mechanisms (e.g. dual-handed motion controllers in virtual reality media) presents yet more challenges in achieving an interactive soundtrack that is both aesthetically pleasing and responsive to player interactions.

A more technical overview of architectures and design processes involved in CGA (particularly practical synthesis) will be presented in the next chapter, alongside a proposed extension to *timbre-led sound models*. Meanwhile, the next two sections serve as an overview of two fields that are central to this thesis: aesthetic and perceptual concepts underlying sound design in narrative contexts, and real-time human interaction with CGA.

[6] http://www.audiokinetic.com
[7] http://www.fmod.com

## 2.3 Sound Design Aesthetics and Frameworks

This section presents a broad overview of some of the fundamental concepts in film sound theory and aesthetic frameworks extending to acousmatic music and sonic art. Film sound theory is a very large field and an exhaustive overview is beyond the scope of this document. Instead some fundamental concepts will be introduced and discussed that will later be referred back to in the broader context of computationally generated audio. More expansive overviews of film sound theory and sound design can be found in Weis and Belton (1985) and Chion (1994).

### 2.3.1 Rendering and Realism

Randy Thom points out the common misconception that good sound design is not about 'creating great sounds' but about enhancing the image (Thom, 1999). In other words, sound need not draw attention to itself but should instead work in collaboration with the visual image. In an interview, Walter Murch states how spectators are more likely to perceive intentional qualities of the accompanying sound in the image rather than in the soundtrack itself, referring to this phenomenon as 'mysterious perceptual alchemy' (LoBrutto, 1994). Indeed the differences between aural and visual perception are well studied in the field of psychophysics. In (Kubovy, 1988) it is argued that visual perception has a spatial foundation and is not inherently temporal, while sound is more intimately tied to temporal perception. This is the foundation of what is more commonly referred to as the ventriloquist effect (Kubovy and Schutz, 2010), where the perceived spatial origin of a sound can be manipulated by a synchronous image. Conversely, sound can alter the temporal perception of visual stimuli as shown in (Vroomen and de Gelder, 2004). In a film sound context, Chion uses the term *audiovisual synchresis* to refer to these effects (Chion, 1994). The pertinent point raised in his definition of synchresis is not the means by which it is achieved but rather the opportunities for generating *added value* (where the summation of sound and image is 'greater than the sum of its parts') through contrasting auditory and visual stimuli. Therefore, for Chion, the purpose of combining sound with moving images is not to reproduce an objectively accurate representation of reality but to *render* 'the feelings associated with the situation'. This, he suggests, results in a more effective and truthful audiovisual image for the spectator. Through the effect of synchresis, spectators have a high tolerance for deviations between sound and image, and this can be exploited with varying degrees of subtlety.

### 2.3.2 Materiality and Embodiment

Chion introduces a further concept of *Materializing Sound Indices* (MSIs) to describe the degree to which a sound references a sense of *materiality* in its source. Materiality

here refers to emphasised features of a sound that cause viewers or listeners to contemplate the material properties of the source, instead of or in parallel to other functions of the source. For example, the pronounced presence of breathing and vocal breaks in a monologue can affect the conveyed meaning of the scene, directing attention to the physical or emotional state of the character in addition to the words themselves. The employment of discontinuous creaks on hollow wood might be more suitable for a tense scene in which a character is walking across an unstable bridge (emphasising the fragility of the structure) than for footsteps on a wooden floor during an important sequence of dialogue (where the materiality of the floor is less consequential).

Parallels have commonly been drawn between the materiality of a sound and the ways in which that sound is perceived in a narrative or musical context. In his proposed 'psychospatial model' of *sound spheres* Sonnenschein (2011) suggests everyday relationships of sound to one's own body and the environment as a way to study the structure of a soundtrack. In the example of the breaking or breathy voice, the spectator can relate to their own sound producing apparatus, thus allowing the scene to evoke meaning on a very intimate or personal level. On the other hand, rendering the same sequence of dialogue with a less pronounced materiality (e.g. using more distant microphone placement) might evoke a more distant relationship to the sound source; more akin to hearing another person speak than relating to the act of speaking itself.

Similar frameworks relating personal proximity to sound materiality can be found in the field of acousmatic music. Young (1996) Young distinguishes between *interior* and *exterior* worlds represented by sound. The latter refers to sounds with clearly perceivable sources while the former refers to variably abstracted sounds constituting a subjective, imagined space for the listener. A similar concept is proposed by Smalley (1997), who refers to the identification of sources in sounds in acousmatic music as *source bonding*.

### 2.3.3 The Source-Sound Relationship

While acousmatic music is a very specialised and, by definition, a sound-only medium, the concept of a sound being able to have a strongly or weakly defined source is a powerful one that warrants further discussion in the context of audiovisual images. As discussed above, onomatopoetic words and vocal imitations are a common occurrence in everyday interactions, and have been observed in children's interactions with their environment from a very young age (Dumaurier et al., 1982). Rocchesso et al. (2015) draw a parallel between drawing and so-called *vocal sketching*, suggesting that information about a sound can be transmitted much more efficiently through vocalisation than through verbal description. Chion relates these sorts of vocalisations to sound design approaches for animations, where sound often closely follows actions depicted

on screen in an approach that is often referred to as *mickeymousing* (Chion, 1994). While this term commonly has derogatory connotations due to the way it imitates on-screen movement on a one-to-one (or *redundant*) basis (Jacobs, 2015), Chion points out that it can have an important functional role including aiding the eye in following complex visual movement (e.g. in animations). As explored in Section 2.1, sound can carry a lot of meaning without having a clear source-sound relationship: 'what is being imitated here is the trajectory and not the sound of the trajectory' Chion (1994). For example, in a vocal imitation of a plane passing by, the pertinent meaning is more likely to reside in perceived qualities of its speed and proximity (conveyed through the temporal and timbral trajectories of the imitated sound) rather than material or geometric properties of the plane (which would be more easily conveyed through a short recording).

### 2.3.4 Temporal Structuring

Many film sound theorists claim that one of the most important attributes of sound is its ability to structure the moving image in time. Out of resistance to redundant techniques (see Section 2.3.3 above) applied in some of the earliest sound films, many film directors attest to the use of so-called *asynchronism* to allow sound to carry a crucial narrative function in counterpoint with the moving image (Jacobs, 2015). In other words, sounds can render sources and actions that are not depicted visually but nonetheless carry an important role in the narrative (for example, the presence of an ambulance does not need to be represented visually when the sound of an approaching siren can be used instead) (Pudovkin, 1985). These techniques have been explored extensively both in film and in writing by influential film makers such as Eisenstein (Eisenstein, 1987) and Tarkovsky (Tarkovsky and Hunter-Blair, 1987), and a detailed overview of the subject can be found in (Jacobs, 2015).

Chion (1994) uses the concept of *temporal vectorisation* to refer to the way sound and image can co-exist in an asynchronous relationship in order to coincide at a particular *synch point*. Thus, asynchronism between sound and image can function as a sort of punctuation of the narrative, by emphasising salient movements (vectors) and moments of reconciliation. While the role of the overall soundtrack on the broader level of film rhythm falls outside of the scope of the work presented in this thesis, temporal structuring of the image through sound can also occur on more microscopic levels, including on the basis of a single sound source. This is a central focus of Chapters 5 and 6.

### 2.3.5 Action-Source Relationship

Complimenting the aforementioned *source-sound relationship*, one can also conceive of an *action-source relationship* to refer to the way in which the sonic representation of a source transmits information about an action. For example, the presence of a ceramic plate can be rendered by sonifying an impact with corresponding material properties – the sound of the same plate fracturing, on the other hand, might direct the listener's attention to the action of somebody throwing the plate against a wall. This concept has frequently been explored in the field of acousmatic music, e.g. in (Wishart, 1996), (Schaeffer et al., 2012) and (Young, 1996).

Hug (2008) distinguishes between listening to a single isolated sound event (and the sonic properties that make up that event) and listening to a soundscape (or sounds in the context of its relationship to a specific environment). The latter could be said to be more relevant to understanding the action-source relationship, while the former relates to the source-sound relationship.

### 2.3.6 Listening Modes

The desire to establish a framework of different listening modes initially arose from the musical use of everyday sounds in a radiophonic context. Through his musical practice Pierre Schaeffer, widely regarded as a pioneer of radiophonic music, realised various ways by which the everyday perception of sound can be altered when manipulating recorded sound. Most fundamentally (as described above) the relationship to the original source can be obscured by the simple act of isolating recorded sound from its visual context. Furthermore, by repeatedly playing back short fragments of recorded sound on a *closed groove*, Schaeffer noted that listeners could be drawn to perceive abstract qualities in the evolution of a sound, regardless of their ability to identify the sound's source (Schaeffer et al., 2012). Another early experiment, referred to as the *cut bell* involved isolating and smoothing the resonant fragment of a bell being struck to obtain a flute-like sound (Chion, 1983). Schaeffer conceived of the term *reduced listening* to refer to this mode of listening, in an endeavour to make a musical formalisation that transcended traditional frameworks based on pitch intervals, rhythm and timbre. In turn, this gave rise to the *sound object* - which refers not to the object that originally caused the sound, but rather a perceptually demarcated sound event that itself can consist of smaller sound objects (Chion, 1983). Schaeffer also conceived of four 'ordinary' listening modes: *listening (écouter)*, *perceiving (ouïr)*, *hearing (entendre)* and *comprehending (comprendre)*, laid out in a matrix format along axes of objective/subjective and abstract/concrete as illustrated in Table 2.2.

Chion later proposed a simplified framework that included two 'ordinary' listening

| | Abstract | Concrete |
|---|---|---|
| Objective | **Comprehending** | **Listening** |
| Subjective | **Hearing** | **Perceiving** |

Table 2.2: Pierre Schaeffer's Matrix of Listening Modes as described by Chion (1983)

modes alongside reduced listening. While *Semantic Listening* refers to the retrieval of meaning through linguistic relationships in the heard sound, *Causal Listening* encompasses all forms of listening based on identifying the sound's cause or its properties. Chion writes: 'causal listening to a voice is to listening to it semantically as perception of the handwriting of a written text is to reading it' Chion (1994).

A popular current perspective takes its roots in ecological perception, where sounds are classified based on so-called *invariant properties* (Gaver, 1993; Clarke, 2005). These refer to fixed qualites perceived in a sound, which conversely facilitate the perception of dynamic behaviours in the sound.

### 2.3.7 Parallels to Music

On a practical level there is some overlap between the concepts introduced above and musical practice (in the traditional acoustic sense). Acousmatic music involves handling and recording of sound-producing objects in the studio (e.g. (Wishart, 1996)). Barrett (2010) explores physical interactions between real sources through intricate spatial projections and manipulations. The relationship between Foley and acousmatic music is explored more explicitly in a project entitled 'Acousmatic Foley' by Pinheiro [8]. Musical recontextualisation of non-musical sound has long been a part of free improvisation and noise music.In addition, instruments are often treated as sound-emulating objects, exploring the boundaries between interior (abstract) and exterior (concrete) sound worlds as described in Young (1996). James MacDonald was originally a percussionist working on animation soundtracks before creating his instrument-like contraptions to perform sound effects (Ament, 2009). In a similar vein, Russolo's *Intonarumori* were built to emulate everyday sounds and employ them, albeit controversially at the time, in a musical context (Russolo, 1913; Chessa, 2012).

In all of these examples a parallel can be drawn to the source-sound and action-source relationships explored above. While in a musical context the materiality of the sound is applied as an augmentation to more traditionally musical ideas, conventional sound design can employ musical trajectories to heighten and punctuate the audiovisual image.

[8] `http://www.sarapinheiro.com`

## 2.4 Real-Time Human Interaction with CGA

In the previous sections of this chapter some historical devices for performing sound effects and environmental sound were introduced, including Luigi Russolo's *Intonarumori* and Jimmy MacDonald's contraptions. This section presents an overview of digital counterparts that have been developed more recently with the increasing ability to use physical sensing devices to interact with computational sound models.

Real-time interaction with computationally generated environmental sounds and sound effects has been a relatively unexplored field until the emergence of Sonic Interaction Design (SID). Much more research has focused on musical interaction with sound, particularly in the design of Digital Musical Instruments (DMIs).

### 2.4.1 Mapping Layers

A significant development in DMI research was the conception of the *mapping layer*, an intermediary computational mediation layer that transforms control dimensions into separate streams of synthesis parameters. In (Rovan et al., 1997) mapping strategies are classified into three fundamental groups of *one-to-one*, *divergent* and *convergent* mappings (the latter two later formalised as *one-to-many* and *many-to-one* (Hunt and Wanderley, 2003)). In a one-to-one mapping, each synthesis parameter is controlled by an independent physical input parameter. A useful analogy to this is a mixing desk, where each potentiometer controls a separate gain value. One-to-one mappings become impractical when dealing with large parameter spaces (due to cognitive load (Tubb and Dixon, 2014)), or when there are fewer physical input parameters than synthesis parameters. One-to-many mappings map a single control parameter to several synthesis parameters and many-to-one mappings use more than one control parameter to control a single synthesis parameter.

Most acoustic musical instruments would fall into the latter two categories (*one-to-many* and *many-to-one*). For example, the pitch of a trombone is controlled by the performer's embouchure as well as the position of the slide and so can be understood as a many-to-one mapping. Variation of energy does not only change the overall volume but also affects the sound quality, implying a one-to-many system. In addition to scaling and other linear transformations of control data within the mapping layer, Rovan et al also suggest the use of more elaborate transformations such as table-lookups and hysteresis to mimic some of the non-linear behaviour observable in acoustic instruments. For example, overblowing the reed of a clarinet results in a drop in amplitude, therefore a simple mapping layer for a wind controller that mimics this behaviour involves a bell-shaped lookup-table (Rovan et al., 1997). Elaborating further on this concept, Hunt and Wanderley (2003) and Wanderley and Depalle (2004) suggest the employment of multiple mediation stages; specifically, one that defines a

space relating to gestures on a physical interface (e.g. 'Sax Lift', 'Energy') and another that maps perceptually meaningful features (e.g. 'Brightness') to parameters of the sound synthesis engine. A similar concept of abstraction layers between physical controllers and sound models is central to the design framework for *timbre-led models* conceived as a central part of this research project and discussed in Chapter 3.

### 2.4.2 Machine Learning Approaches

Other work focuses on the development of generic and modular mediation strategies that can exist between an arbitrary pairing of physical controllers and synthesis engines. One such approach is the implementation of machine learning procedures in order to learn mappings between control and synthesis parameters. In possibly the first proposed application of *neural networks* in a DMI context, Lee and Wessel (1992) mention the mitigation of long learning curves and motor skill adaptation (by having the machine adapt to interaction rather than the opposite way around) as a primary motivation. (Bevilacqua et al., 2009) introduce the concept of *following gestures*, using Dynamic Time Warping and Hidden Markov Models to recognize trained gestures and precisely follow their execution and variation over time. This work is extended further by Francoise et al (Francoise et al., 2012) to include real-time segmentation of gestures (following Godøy et al. (2009)'s theoretical frameworks of *chunking* or *coarticulation*), enabling more precise control over sound parameters using abstract gestures that are meaningful to the given user or designer. These techniques are commonly referred to as *mapping-by-demonstration*, denoting the process of iteratively performing gestures in synchrony with sound output to train the computational mediation layer. In (Fiebrink et al., 2009) a more general purpose tool is presented, allowing for so-called *on-the-fly* machine learning, wherein players can performatively develop mapping layers. A detailed overview of machine learning techniques in interactive musical contexts can be found in (Fiebrink and Caramiaux, 2017).

### 2.4.3 Physics-Based Mediation

Momeni and Henry (2006) propose the use of an independent physics-based control layer to control audio and/or video synthesis. This is achieved by actuating simulated physical systems such as mass-spring models using gestural input and connecting the output of the control layer to synthesis parameters in some meaningful way. Although no particular type of synthesis is specified here, this approach can also be applied to the control of physical or physically inspired sound models. Such an implementation was carried out recently by Thoret et al. (2013) using a model of a bowed string as a proxy for generating parameters of a friction-based sound model. A similar control strategy was developed in this research project (referred to as a *Physically Inspired*

*Control Layer*) and is described in Chapter 4 and in (Heinrichs et al., 2014; Heinrichs and McPherson, 2014)

Gohlke et al. (2011) suggest various approaches to mapping accelerometer-based controllers to environmental sounds as ways of performing 'virtual Foley'. An example of such a mapping strategy is 'shake mode', where 'aggregate motion of the device is mapped to the intensity and rate of repetition of sound events'. Among the suggested sounds to control in this way are not just the ones that are closest to the action - for example, a jar full of coins - but also other sounds that share the aforementioned concepts of 'aggregate motion', 'intensity' and 'rate of repetition' - i.e. cheering audiences, steam engines and swarms of bees.

In a study exploring expressive controllers for physical models of bowed strings Serafin et al. (2001) propose the development of hybrid controllers. Using the expressive affordances of a saxophone, including the haptic feedback and the complexity of key-fingerings, a separate set of expressive parameters are controlled, including bow pressure, bow force, centre frequency and frictional properties of a physical bowed string model Burtner (2003).

Smyth and Smith III (2002) apply the concept of physics-based mediation on a mechanical level (instead of a computational abstraction), whereby the so-called *buckling* mechanism observed in Cicadas is imitated by a series of aluminum ribs. The dynamic properties of the interaction result in a natural controller-synthesis pairing due to force feedback intrinsic to the device.

### 2.4.4 Ergotic Interaction and the Enactive Approach

The above implementations are usually based on physical models, where virtual representations of mechanical movement provide both sonic and haptic feedback channels. This approach to interaction with digital representations of objects is commonly referred to as *ergotic*, first proposed under the broader framework of *instrumental gesture* (Cadoz, 1988). Ergotic interaction has been claimed to be central to the *enactive* approach to musical interaction (a term borrowed from the field of tangible human-computer interaction) (Cadoz, 2009), where tacit gesture-object relationships learned from interactions in everyday environments are exploited to achieve natural sonic interactions (Essl and O'Modhrain, 2006; Armstrong, 2006). In the enactive approach to sonic interaction a physically realistic simulation of sonic objects and mechanical feedback is not strictly required, but instead a sufficiently *close* link between action and perception is encouraged by means of *energetic consistency*. An interesting anecdote is provided by (Hunt et al., 2003), who describe how an attempt at recreating a Theremin resulted in an error whereby the velocity of hand movement controlled the oscillator's amplitude, rather than the hand's absolute position. This resulted in a more engaging and nuanced interaction where sound is only outputted during

moments of movement: in order to produce sound output, a more-or-less consistent amount of energy needs to be fed into the system through physical interaction.

### 2.4.5 Guiding Interaction through Sound

Enactive approaches to sonic interaction are particularly relevant to the control of physics-based models of everyday sounds. Here the focus shifts towards the ecological affordances of sound models (in other words, the actions that are implied by the represented sound object). Caramiaux et al. (2011) distinguish between *causal* and *non-causal* sound in a user study suggesting that causal sounds are more likely to elicit gestures corresponding to the sound's perceived causation. Interactional frameworks based on the causality of sound (as opposed to timbral or musical properties) are particularly relevant in the fields of Sonic Interaction Design and Interactive Sonification. In (Delle Monache et al., 2008a) interactions with commodities are augmented with contrasting sound models of everyday sounds in order to guide interactions. For example, a bowed-string model is used to help identify the ideal tightness at which to assemble a coffee moka. *Model-based sonification* exploits the perceived physical properties of virtual sonic objects to facilitate the interactive exploration of a data-set (Hermann and Ritter, 1999) - for example shaking a bottle in order to approximate a percentage value based on the perceived amount of contained liquid or particles.

### 2.4.6 Criteria and Evaluation Metrics

Despite the large body of work carried out in the field of DMI design and other fields of sonic interaction, there has been comparatively little focus on the development of evaluation strategies. Wanderley and Orio (2002) propose the application of metrics from the domain of Human-Computer Interaction (HCI) in the evaluation of musical interfaces.Examples of such metrics include *repeatiability* (the ability to successfully repeat a given sound) and *navigability* (ease of navigating the instrument's parameter space). The application of HCI methods has been criticised for ignoring issues central to musical expressivity in favour of a task-oriented evaluation of a given interface. In a reflective evaluation of HCI methods, Kiefer et al. (2008) point out how such tools insufficient to account for experiential qualities of the interaction and suggest incorporating affective qualities (e.g. physiological data) to enhance and contextualise data. Stowell et al. (2009) suggest that HCI approaches can be useful when applied to isolated tasks that are known to contribute to the experience of musical expressivity. More formal qualitative methodologies such as discourse analysis are recommended in the evaluation of complex musical interactions. O'Modhrain (2011) calls for evaluative strategies to take into account various *stakeholders*, who may have distinct criteria in the assessment of a DMI. For example, what a performer considers to be an expressive

instrument may not be perceived as such by an audience watching a corresponding performance.

Drawing comparisons to traditional (acoustic) musical instruments Jordà (2005) outlines several aspects that should be taken into account in the design and evaluation of DMIs. Three levels of a musical instrument's *output diversity* are proposed as playing an integral role in contributing to its potential for expressivity. *Micro-diversity* refers to the ability to vary nuances of the instrument's output, thus enabling a performer to give their own interpretation of piece of music. *Mid-diversity*, on the other hand, is a measure of how much a given instrument allows two pieces of music to differ from one another. For example, a percussive instrument with no determinate pitch, such as a snare drum, may be considered to have a lower level of mid-diversity than a xylophone to some listeners, but a higher level of micro-diversity due to nuances in timbre and dynamics. *Macro-diversity* is a measure of an instrument's flexibility to be applied in different musical contexts; for example, a guitar might be seen to have a larger level of macro-diversity than a piccolo. These concepts will be revisited in Section 4.4.4.

Another common theme that has attracted a lot of focus in last decade of DMI research concerns virtuosity of the instrument, and the necessity of overcoming a learning curve in the learning process (Dobrian and Koppelman, 2006; Jordà, 2005). This points to longitudinal studies (O'Modhrain, 2011) in order to account for instruments that, for example, despite having a 'low entry-fee' (are easy to adopt) should ideally also have a 'high ceiling' (accommodate the development of virtuosity over much longer time periods) (Jordà, 2005).

### 2.4.7 Sonic Interaction in Games

A fundamentally interactive medium, sound has been shown to fulfill a number of different roles in games, including feedback pertaining to game states (e.g. winning or losing), projected affect (e.g. eliciting fear in the player) or simply emphasising the presence of an object or an environmental setting. A detailed overview is beyond the scope of this thesis but can be found in (Wilhelmsson and Wallén, 2010) and (Collins, 2013). Pertinent to the above discussion of direct interaction with sonic objects is the use of *ego-centric sound* in games. Ego-centric sound or ergo-audition is a term borrowed from (Chion, 1994), which describes all first-person interactions in a game usually represented by an avatar (i.e. virtual representation of the embodied player-character). Unlike musical instruments and experimental sonic interfaces developed in research labs, games are interacted with by means of extremely standardised control mechanisms, which up until recently have consisted of sets of binary buttons and one or two 'joysticks' providing two analog axes of control. The function of these controllers is of course not to directly perform sounds or music, but to interact with a

virtual environment. Thus sound accompanying actions such as 'jumping' or 'shooting' (triggered by the press of a button) is usually designed to facilitate a high level of immersion and engagement. The low interactional bandwidth means that instead of designing a dedicated synthesis engine, waveforms can simply be played back on a per-event basis, allowing traditional approaches from film to be applied. Nonetheless, sound designers commonly draw parallels between this aspect of game sound design and musical instruments. In an interview in the game audio documentary 'Beep', established game sound designer George Sangers states: "I always felt like i was writing for a very interesting group of one-armed musicians [that can produce] two boops, a beep and a pfft"[9].

### 2.4.8 Ergo-audition and Performed Sound

Caramiaux et al. (2011) and Godøy (2010) distinguish types of gestures that mimic a sound's cause from those that trace its inherent morphology. While both types of gestures could be said to be 'performative' the latter type is likely to be more suitable in creating complex sound trajectories as part of a design task, as it provides an easier means for controlling the sound along abstract timbral dimensions.

Hug acknowledges the possibility of employing performative strategies in the design of interactive sonic objects (Hug, 2010). He refers to the mediated nature of immersive environments as a 'second order' experience, noting an element of re-enactment when interacting with sounds in a virtual environment (i.e. ergo-audition through the actions of an embodied avatar). This affords a different set of design strategies from those employed in sonically guided interactions (for example in the *sonically augmented found objects* presented by Delle Monache et al. (2008b)), where the focus is on enhancing interactions with everyday objects (or 'commodities') through the use of sound, and not on the imitation or design of sound.

### 2.4.9 Gesture and Sound

The term *gesture* is used widely in the field of sound and music computing and can take on a variety of different meanings (Jensenius, 2014). As such it's worth briefly exploring this term in the context of sound design. Smalley defines gesture as an "energy-motion trajectory which excites the sounding body creating spectromorphological life" (Smalley, 1997) noting that the process also works in reverse, i.e. the listener perceives human expression through gestural activity within the morphology of sounds. A scientific basis for these claims is provided more recently in (Godøy et al., 2009) and (Godøy, 2010), who suggest that humans possess a highly sophisticated mechanism for decoding gestures from sounds based on hierarchical

[9] http://www.gamessound.com

segmentations of actions. The degree of detail in gestural information that can be extracted from sound signals is exemplified in recent work by Thoret et al. (2013), which investigates the use of computational models to guide gestures along two-dimensional surfaces. This supports the claim that the extensive use of Foley art in film is not only due to its cost-efficiency in particular scenes that contain a lot of movement, but also because it is a performance in itself that adds vital 'life' to the image (Ament, 2009).[10]

Worth noting here is that unlike the concept of performing timbral spaces introduced above, Foley corresponds more closely to the reenactment of action-sound relationships by means of manually handling real objects. Thus performed behaviour in the context of Foley art corresponds explicitly to a broader scale of actions and events rather than the design of sound effects on a timbral level. In this thesis these two scales of performed behaviour correspond to the two phases of design and integration of computational sound models proposed here. While designing models involves consideration of timbral and perceptual features to compose expressive deviations from physical behaviour, the performance of action-sound behaviours in synchrony with continuous movement is more relevant to the problem of integrating sound models back into an interactive environment.

### 2.4.10 Interacting with Environmental Sound

Controllers and interaction strategies for environmental sound have been the focus of several research projects in recent years. The majority of these focus on augmenting natural interactions with corresponding or closely related sounds. In (Essl and O'Modhrain, 2005) an interface resembling a whiteboard eraser was used to control friction and other sounds produced by granular and wavetable synthesis. Here spectral features were extracted in real-time from microphones mounted on the surface of the interface allowing players to use scrubbing actions to control new sounds. Böttcher and Serafin (2009) used a commercial 3-axis motion controller (Nintendo Wii) to sonify sword strokes in a 3-D game environment. Müller-Tomfelde and Münch (2001) and Chung et al. (2013) investigated ways of sonifying pen sounds to augment the experience of using digital pen and surface controllers. Serafin et al. (2006) recreated Russolo's Intonarumori (introduced in Section 2.1) using a novel physical hardware controller and a dedicated synthesis engine. This implementation is particularly interesting as it combines physically informed source models and an enactive interface to recreate *approximations* of environmental sound, as originally envisioned by Russolo (Russolo, 1913).

[10] This is explored more thoroughly in Chapter 5 in reference to a questionnaire conducted with Foley artists

More recent work conducted as part of the *SKAT-VG* project[11] focuses on the use of the voice to perform or *sketch* sound effects using real-time synthesis models. Houix et al. (2016) present a toolkit for controlling parameterised sound models, such as car engines, using features extracted from real-time microphone input. This allowed the performance of sound effects by imitating the targeted sound behaviours using the voice. This is explored in the context of *sonic sketching*, where sound designers rapidly explore sound behaviours using their voice, without the need to use microphones, physical sources or sound libraries. While timbral limitations precluded the employment of vocal sketching in professional sound design practice, this could potentially be mitigated by the coupling of real-time synthesis with vocal imitation (Baldan et al., 2016). The consideration of vocal sketching in the control of sound models was also studied in a social and pedagogical context is explored by Ekman and Rinott (2010).

## 2.5 Discussion

**Realism and Sonic Impressionism**

Musical scores by Gerard Grisey and Tristan Murail from the French *Spectralism* movement were composed for orchestras to carefully emulate real-world sounds with the aid of state-of-the-art analysis and resynthesis methods and collaborations with psychoacousticians (Pressnitzer and McAdams, 2000). Through their musical practice they developed a language and aesthetic based on sound as 'pure and abstract materiality' (Anderson, 2000). Though the ability of audience members to 'listen beyond the orchestra' can and has been contested by the composers themselves (much like Schaeffer's concept of *reduced listening*) (Murail, 2000) the concept of composing sound objects like paint on a canvas (Malherbe et al., 2000) is a powerful one that has only recently started to be explored in the domain of computationally generated audio.

Puronas (2014) draws this comparison between painting and sound design for the moving image in a discussion of so-called *sonic hyperrealism*. Figurative painting channels meaning through material qualities such as brush strokes, while maintaining a reference to the depicted source. Puronas suggests taking a similar approach to sound design, where through the process of imitation with a specific focus on timbral qualities of the sound, one can imagine an *impressionistic* aesthetic emerging. This is put in contrast to the dependency on recorded sound, which is likened to a process of 'sound montage' or 'sonic taxidermy'.

A similar argument is made by Farnell (2011), who distinguishes between *sensible* and *essential* realism. In the former case, recorded sound features realistic qualities

[11] http://skatvg.iuav.it

on a superficial level ('surface reality') but is unable to portray realistic *behavioural* qualities. Essential realism, instead, is central to Farnell's process-oriented approach. While surface features of the sound are likely to be less true to reality than a recording, the way in which the sound behaves and reacts in an interactive setting portrays a greater sense of realism (in the sense of liveness) that is impossible to obtain with recorded sound.

This argument can be taken further to criticise approaches within the domain of CGA itself. In a reflection on the transdisciplinary *Sounding Object Project* (Rocchesso and Fontana, 2003) Vicario (2003) questions the suitability of physical models in recreating audible behaviours for virtual sound sources. A persistent focus on physical simulation results in sterile representations of 'perfect' geometries that would never be found in nature. Vicario argues that a perceptually-led approach would instead be more suitable, drawing comparisons to techniques in Foley, where sound sources are commonly rendered using unrelated objects (e.g. umbrellas for birds' wings, twigs for breaking bones).

The concept of realism is central to the discussion of both computational sound models and virtual environments. From a technical point of view, realism is regarded as the end-goal in physics-based approaches. On an objective level this is a reasonable perspective, but in a perceptual context realism has been shown to negate objective ideals (Mengual et al., 2016) (Lennox, 2004). There has been less focus on technical development that supports expressivity in the audio-visual relationship as discussed in Section 2.3.

### Revisiting *Expressivity*

If in musical instruments the concept of *expressivity* is typically associated with control intimacy, in games it would be more easily associated with the sonic enhancement of an interaction with the virtual environment by means of exaggeration of the accompanying sound's causal and/or behavioural properties. With the recent reemergence of virtual reality in mass-market entertainment and, by extension, the employment of continuous controllers with huge interactional bandwidth (e.g. more than six degrees of freedom per hand), designing sound that satisfies both traditional criteria of sound design while providing adequate feedback to continuous interaction remains an open question. As discussed above, current sample-based approaches are limited in their ability to provide appropriate feedback to such rich control data, but physics-based approaches lack the aesthetic flexibility of conventional approaches. This thesis aims to address this problem by exploring new sound design procedures based around human performance. In other words, one form of human interaction (performance) is integrated into the design process in order to facilitate more stylistically rich end-user interaction with virtual sounding objects.

A largely overlooked flaw in the current state of CGA is that there is little means of adding expressive nuances to computationally generated sounds. Previous literature has focused largely on computational audio as a means of heightening interaction with everyday objects (Rocchesso and Fontana, 2003; Delle Monache et al., 2008b; Böttcher and Serafin, 2009), but less focus has been dedicated to the nuanced control over timbral qualities in translations of physical interactions to sound. Stylising a sound while preserving its ecological qualities is a challenging research problem that is central to this thesis. This is addressed from ecological and pedagogical standpoints by (Hug, 2010) and (Altavilla et al., 2013), using prototyping techniques including live Foley and vocalisation (Piccolo and Rocchesso, 2016).

This research project, on the other hand, investigates this on a more pragmatic level from the point of view of:

- Synthesis architectures (i.e. how to design physically-inspired models within a timbre space using performed gestures)
- Integration techniques in a game audio context, with a focus on techniques applied by Foley artists in non-interactive media.

Chapters 3 and 4 are centred around the first of these two points, examining design processes underlying practical synthesis models and exploring ways of interacting with such models to generate evocative sound effects. Chapters 5 and 6 focus on the integration of sound models into an interactive environment, juxtaposing current approaches of mapping computational sound to images with performative approaches taken by Foley artists. Finally, findings from these studies are discussed in the broader context of CGA, sound design and human interaction in Chapter 7.

# Chapter 3

# Extending Practical Synthesis

Computationally generated audio provides degrees of responsiveness that are difficult or impossible to achieve with recorded waveforms. While numerical approaches are capable of producing physically realistic behaviour in response to a tangible parameter space, they impede opportunities to creatively intervene in the perceptual features of the resulting sound. Aesthetic shortcomings due to design processes focused on realism have only recently started to be addressed, particularly in the work of Farnell (2011). This chapter provides an overview of practical synthesis techniques proposed by Farnell, with a particular focus on the underlying design strategies. Central to this is the concept of *behaviour* in sound, which will be defined within the context of a practical model. While numerical models intrinsically simulate dynamic behaviour pertaining to the modelled source, practical models allow a more clear separation between behavioural and nested signal-generating components.

A new category of *timbre-led models* is proposed, wherein the behavioural components of the model are omitted in place of a subjective parameter space pertaining to perceptual sonic features identified in the target source. By exposing a perceptual space (in favour of a physics-based behavioural space), behaviours can be authored externally to the timbre-led model, potentially facilitating opportunities for stylisation and expressivity that would be harder to obtain in a procedural behavioural model. This forms the basis of a more focused investigation in Chapter 4, applying timbre-led approaches in both the design and real-time performance of sound effects.

Section 3.1 serves as an overview of Farnell's concept of behavioural audio, contextualised within his proposed design strategies for practical synthesis. Section 3.2 proposes the novel sub-class of timbre-led models based on an extension to Farnell's design strategy.

## 3.1 Behaviour in Practical Synthesis Models

### 3.1.1 Behavioural Audio

Central to the design principles of practical synthesis set out by Farnell (2008) is the concept of *behavioural audio*. Farnell defines this in contrast to the use of static samples, where in the domain of visual media an analogy can be made to a photograph. A recording, just like a photograph, exhibits what he refers to as *sensible realism*, wherein only surface features of the sound are defined, but are enough to render its behaviour in a very narrowly defined context. It functions like a static audible snapshot of an object's behaviour, usually with a degree of detail and complexity that is difficult to simulate by computational means. However, Farnell notes that it lacks in so-called *essential realism* in that it is impossible to interpolate between two such snapshots in a physically or perceptually consistent way because there is no underlying *behavioural model*. For example, given two recordings of a bottle being filled with water at different speeds, it is difficult to produce an intermediate sound without an underlying model of how different physical states affect the observed behaviour. Instead, a computational model would incorporate dynamic properties such as volume of flow, surface area and air cavity size to facilitate a wide range of behaviours associated with the speed of pouring.

A behavioural model then is something that outputs perceptually consistent (consistency used here in favour of 'realism' for reasons described below) sounds as a result of manipulating a set of exposed variables. The object that is being designed is not a concrete (in the Schaefferian sense (Schaeffer et al., 2012)) sound that exhibits a single, albeit potentially well-crafted, behaviour. It is instead a *process*, a system with variable states and properties that can generate a multitude of behaviours in a perceptually consistent way. Thus behaviour, in this sense, can be defined in relation to the parametric space of a computational model: "A well-formed parameter space provides a behaviour captured by the fewest salient variables while allowing the greatest sensible range" (Farnell, 2011).

### 3.1.2 Source Models

The most straightforward way of achieving a parametric space that generates a consistent range of behaviours might be to model the physical properties of the source directly. Source models (or physical models) are based on virtual representations of spatial geometry and materials, simulating wave propagations and other mechanical interactions from first principles. The resulting physical behaviours are consistent with interactions from the everyday environment (though obviously constrained by the scientific accuracy of the model) as an intrinsic property of the model.

While a purely physical-analytical approach to the development of a sound model

is a perfectly valid pursuit, there are various ways in which it can fall short of being a valuable asset. The first is in computational efficiency. Numerical models that approach even modest levels of geometrical complexity can require a large number of compute cycles, with most brute-force methods still being far from the ability to run in a real-time context (for example (Zheng and James, 2011)). Even in the case of a viable real-time implementation, it can lead to a situation where the simulation overcompensates for the *behavioural breadth* that is required by a given context. A good example can be found by making an analogy to Foley art. For example, a Foley artist might use a leather belt to produce the sound of squeaks for a given sequence of movement. The equivalent source model of the leather belt would not only account for behaviours associated with 'squeaking' but also for all other interactions including rolling, whipping and so forth, which would not be required in this context. For reasons of computational efficiency it would be favourable to develop a model that is constrained to the required behaviours, and this may often be difficult to accomplish with purely numerical methods.

A *signal model* is less strict in its implementation and emulates the target source on the basis of both physical components and observed sonic features that can be interpreted as a chain of signal processing blocks upon iterative stages of analysis. This makes it possible to reduce the complexity of a model through a targeted design process, described below. As will be discussed later on, there is also aesthetic merit in this approach, as it opens up more access points for creative intervention in the architecture of the sound model.

### 3.1.3 Schools of Design

Farnell (2011) describes several *schools of design*, that can be followed in the use of computational methods for designing sound. Three of these are worth highlighting here, regarding the integration of behaviour into a computational model: the *essentialist*, the *behavioural* and the *phenomenal* schools of design.

The essentialist school is associated with brute-force implementations, such as Finite-Difference Time Domain and attempts at solving the wave equation for arbitrary spatial configurations of objects. By developing a detailed numerical simulation of a physical source, audible behaviour can directly be extracted (e.g. by measuring values of velocity or air pressure at a particular geometrical point). Behaviouralists are primarily interested in recreating a particular *subset* of behaviours through perceptual or physical simplifications of a more elaborate model. Finally, the phenomenal school accepts any dynamic sound generation technique as long as it can recreate a single or small subset of behaviour.

Perhaps one of the most important contributions by Farnell is in his proposal of design strategies that treat these schools of thought as a continuum (following what

he refers to as the *pragmatic* school). Central to this approach is the application of various analytical processes in order to understand the physical mechanism underlying the source's behaviour, while also being able to make perceptual simplifications or stylistic divergences from reality.

### 3.1.4 Design Strategy

Taking inspiration from software engineering principles, Farnell (2008) suggests a *top-down to bottom-up* approach in the design of computational models, where the model is designed top-down and implemented bottom-up. Having determined the functional and/or aesthetic context of the sound model, the design process starts with the collection of reference materials such as recordings and physical descriptions of the targeted sound source. From these reference sounds, a (theoretical) *model* is constructed that outlines the required *components* and *behaviours* of the object. Going back to the example of pouring water, the structural components might initially include a resonating cavity and a body of water while the behaviour consists of filling the bottle of water (thereby increasing amount of water and shrinking the air cavity) at various speeds. A further stage of analysis might break this model down into a finer grain of detail, for example expressing the body of water in terms of bubbles and surface disturbances that arise as a function of the pouring speed. In the next stage of the process, a set of synthesis *methods* is chosen to simulate components of the abstract model, which can vary from conventional methods such as additive and subtractive synthesis to more elaborate ones such as physical models (e.g. waveguide-based approximations of resonant bodies) and other dynamical processes (e.g. chaotic attractors). Finally, the model is *implemented* in a software environment that is most appropriate for the chosen methods. For example, a model that incorporates many time-dependent dynamical processes might be best implemented in a low-level textual language that expresses signals on a sample-by-sample basis, whereas a predominantly subtractive approach might be more suited to a high-level language or graphical environment based on atomic unit generators (e.g. Puredata or Supercollider).

Following the *bottom-up* paradigm, the process of implementation begins with the technical development of the *signal chain* (in other words, a series of interconnected signal process blocks corresponding to the developed model and methods), followed by further abstraction or encapsulation (in other words simplifying the technical architecture to form a more overseeable high-level representation), leading to the final *parameterisation* of the model. It is in this process of parameterisation and abstraction that the model's behaviour is defined. This will be illustrated below using the simple example of a sword swooshing model.

### 3.1.5 Designing and Implementing Behaviour

A physical analysis of air turbulance around objects reveals that features of the resulting sound are primarily dependent on the geometry of the given objects. While large irregular objects produce sound through resonating arbitrarily sized cavities, long or cylindrical objects are set into quasi-periodic motion, giving rise to air vortices producing so-called *aeolian tones*. In each case the resultant sound can be approximated broadly using a band-limited noise source, where the width and shape of the frequency range corresponds to the size and symmetry of the object (see (Farnell, 2008) for a detailed overview of the model and the corresponding analytical steps).

#### Signal Chain

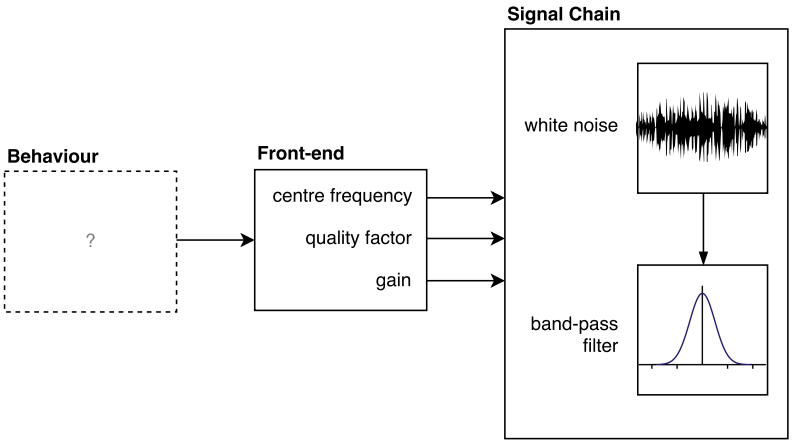


Figure 3.1: A basic model of wind and its parametric front-end.

The resulting signal chain is very simple, consisting of a white noise generator passed through a band-pass filter with variable *centre frequency*, *quality factor* and *gain*, for each wind-obstructing object in a scene. Figure 3.1 shows the basic signal model for each source as implemented in Farnell (2008) and various commercial audio synthesis plug-ins. As discussed in Selfridge et al. (2016) and Farnell (2008), gain and centre frequency of the filter can be shown to correspond to the velocity of air turbulence around an object (as caused by a gust of wind or a swinging sword), while the filter's resonance corresponds to the geometry of the object. These three parameters make up the *front-end* of the core signal chain. While this parameter space is sufficient to recreate the sound of air movement, the values need to be changing dynamically over time in order to exhibit plausible behaviour. Even if set to static values within a meaningful range, the output of the signal chain would bear little resemblance to the sound of wind (e.g. wind never moves at a perfectly constant velocity). Further *be-*

*havioural abstraction* (Farnell, 2008) is required in order to incorporate the temporal behaviour that gives the model its 'windy' or 'swooshy' quality.

### Behavioural Abstraction

The signal chain shown in Figure 3.1 is capable of generating a reasonable approximation of air turbulance for a variety of obstructions, however some additional abstraction is required in order to obtain a parameter space that is consistent with a physical description of the source.

In the early design stages of the model this source component might have been referred to as a 'wind obstruction'. The wind obstruction produces physically viable behaviour as a function of a dynamic *air speed* parameter. Physical analysis has shown that wind velocity affects both the gain of the noise generator and the centre-frequency of the band-pass filter, according to the high-level model. In addition to this dynamic parameter, the model specifies that the obstructing object's size corresponds to the range of the centre-frequency modulation while its symmetry affects the quality factor of the band-pass filter. This transformation of a high-level *behavioural* parameter space (air speed, size, symmetry) into a low-level *signal* parameter space constitutes the behavioural abstraction, effectively turning an abstract signal processing block into a tangible representation of a source.

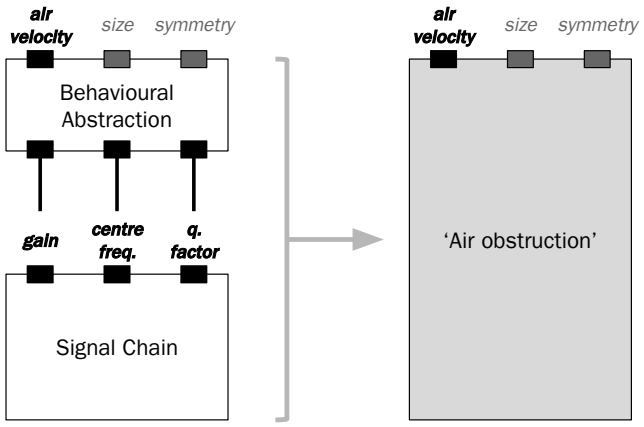


Figure 3.2: Behavioural abstraction to form 'air obstruction' model. Bolded labels correspond to dynamically varying parameters.

Yet further behavioural abstraction can be performed on the new parameter space. For example, Farnell's implementation of a windy scene includes a variety of wind obstructions, each placed at a different point in virtual space. In order to generate physically consistent behaviour, a combination of low frequency noise sources and oscillators are used to generate a pseudo-random signal corresponding to gusts and

squalls of wind. This signal is then passed to each obstruction object with temporal delays corresponding to their positions in space. By scaling this signal by a *windiness* parameter, a new behavioural model of a windy scene is formed.

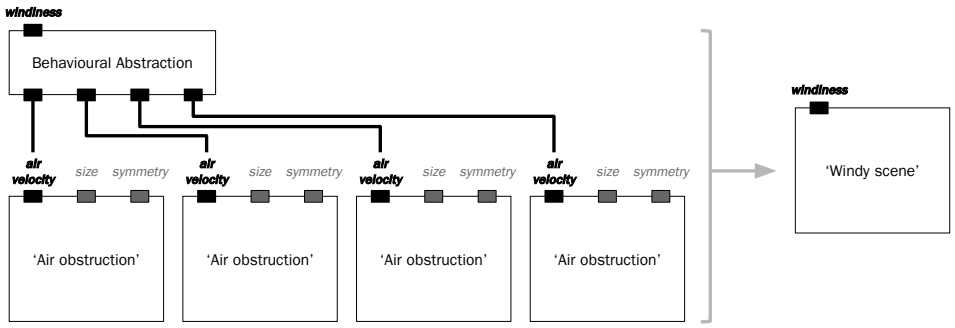


Figure 3.3: Nested behavioural models to form 'windy scene'. Bolded labels correspond to dynamically varying parameters.

Alternatively, the original parameter space of a single obstruction source could be preserved and used to represent a sword. An external parameter corresponding to the *swing velocity* of the sword is used to control the air velocity of the obstruction. Additional parameterisation could be applied to the filter coefficients to simulate the effects of changing the sword's angle.

### 3.1.6 Creative Intervention

Farnell's design process is particularly suitable for creative intervention in an otherwise objectively informed model. In deriving the behavioural parameterisation of the sword model, the designer might decide (at any given stage of analysis) that the model should react more acutely to larger values of air speed, producing higher frequency ranges and more resonant tones. This decision could be informed by an acoustic analysis of cinematic sound effects in the process of developing the high-level model, or as a result of experimentation while implementing the model. The model can no longer be said to be a physically accurate simulation, however it has been tailored to conform more closely to the designer's preferences. These may be purely stylistic in nature or motivated by the targeted function or application of the model.

If numerical modelling is the 'brute force' physical approach to designing computational sound models, these creative interventions can be placed on the opposite end of a spectrum where adherence to stylistic preferences or functional requirements supersede the need to achieve physical accuracy. Exaggerating particular features of a sound effect is a common technique in sound design, also sometimes referred to as *cartoonification* (Lennox and Myatt, 2011) or *caricature* (Back and Des, 1996).

Common examples of such exaggeration include psychoacoustically augmenting the overall spectral qualities of the sound in order to be more perceptually salient to the listener, or combining multiple sound sources to emphasise a particular event (e.g. a door slam). Many of these techniques can be found in early radiophonic and electro-acoustic music, which relied almost exclusively on the editing and superimposing of recorded sound to produce hyperreal impressions of everyday sounds. Computational audio gives rise to new possibilities in this realm by making it possible to control arbitrary aspects of the sound model while keeping its *invariant structure* (Windsor, 1995; Clarke, 2005) - the *signature processes* Farnell (2014a) that define its acoustic verisimilitude - intact.

### 3.1.7 Summary

Design strategies proposed by Farnell offer an attractive alternative to numerical simulation methods from both angles of computational efficiency and reappropriatability. The *top-down to bottom-up* design strategy implicitly makes it possible to separate behaviour from signal processes through the process of parameterisation. This stands in contrast to the design of source models, where audible output is generated as a consequence of a precise but opaque physical simulation (see Figure 3.4).

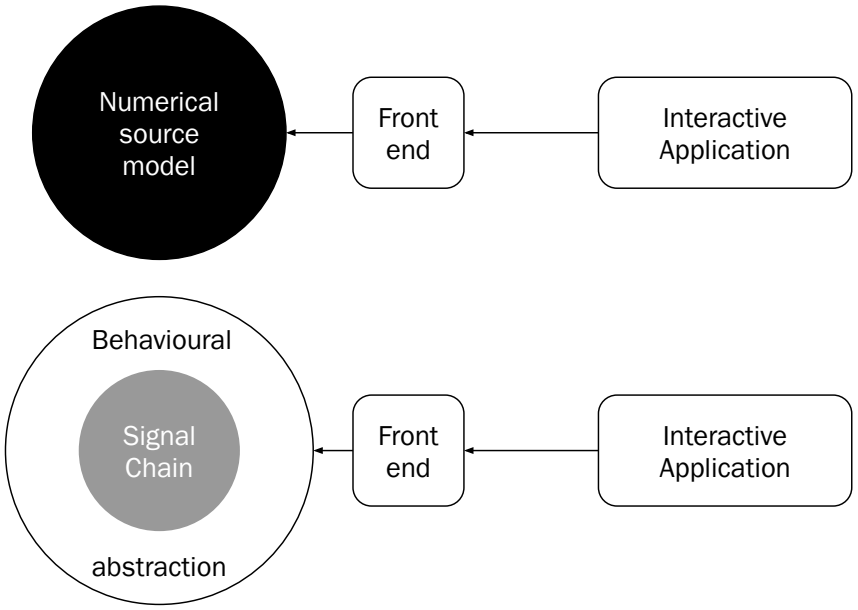


Figure 3.4: Behavioural abstraction in a practical synthesis model and the 'black box' nature of a numerical source model.

The next section introduces the concept of timbre spaces and the development

of models based predominantly on sonic and perceptual features. A new sub-class of *timbre-led models* is proposed, wherein behavioural abstractions are intentionally eliminated to expose a perceptual front-end, rendering the development of behaviours a *compositional* task that is external to the model.

## 3.2 Designing Behaviour in Timbre Spaces

Physically-informed design, as found in the practical synthesis approach, is a reliable strategy for integrating complex and believable behaviour into a sound model. Further to this, parametric simplifications of physical processes in the signal domain raise opportunities for creative intervention. While the behaviour is encoded parametrically based on simplifications obtained through physical, signal or psychoacoustic analysis, the modular architecture allows it to somewhat break free from the constraints of physical laws.

### 3.2.1 Behavioural Timbre Space

In the above example of the air obstruction object three physically viable parameters (one dynamic and two fixed) were used to control three lower-level parameters of the underlying signal chain. The latter three describe a parametric space that encompasses the whole breadth of possible sound morphologies that can be associated with higher level behaviours. By considering this space in a perceptual rather than a physical context one can conceive of a *behavioural timbre space*, where behaviours are defined as trajectories navigating a space of 'intensity', 'pitch' and 'resonance'.

The concept of interpolating across perceptual features takes its roots in a study on *timbre spaces* by Wessel (1979). Subjects were told to identify perceptually contrasting sounds from a set of examples. This was mapped to a continuous space of parameter settings for an additive synthesiser, which made it possible to meaningfully blend across perceptual features.

Following the design process described in the previous section, the resulting signal model is likely to already incorporate perceptual descriptions of parameters when psycho-acoustic and signal analysis is employed alongside a physically-informed approach. In the example of the wind model above, the effect of air cavities in irregular surfaces and air vortices was modelled based on a psycho-acoustic reduction that described the resulting sound in terms of 'frequency ranges', 'loudness' and so forth. In fact, in the implementation described in (Farnell, 2008) the underlying object is labelled as a 'whistling' model, describing the perceptual acoustic qualities of the sound as opposed to the broader physical behaviour.

A hypothesis is stated here wherein by making a clear distinction between behavioural space and timbre space, behaviours can be composed independently of the

model. To illustrate this concept an analogy can be made to the animation of three-dimensional models for games and films. Here the model consists of three-dimensional geometric meshes with prescribed transformation points along which the corresponding part can be moved, scaled or rotated. Regardless of how detailed the model is, it needs to be *rigged* (a parameter space defined) and *animated* in order for it to exhibit any plausible behaviour. Animators can use a simulation engine to generate behaviours based on physical principles, but will often instead define them manually in order to obtain heavily stylised and expressive movement. Unique behaviours are defined for specific actions (e.g. jumping or throwing) and played back and interpolated in real-time to achieve seamless sequences of movement. The parameter space addressed by these behaviours is normally very abstract, referring to object positions and rotations specific to the particular model. Transferring the concepts defined above, it would mean that each animated action is assigned its own unique behavioural layer, transforming a higher-level parameter space into meaningful streams of parameter values driving the model.

A similar approach can be taken to the control of computational sound models. Instead of incorporating computationally formalised (or *procedural*) behaviours, a lower-level parameter space could be exposed as the front-end to the model, referring to idiosyncratic features pertaining to subjective descriptions of sound transformations.This allows unique sequences to be designed, serving as impressions of particular behaviours rather than carefully modelled simulations of objective movement. This way, physical verisimilitude (as far as it is obtainable by the base model) is no longer a limiting factor for stylisation, but a choice made by the designer.

Just as sound designers and composers of electro-acoustic music emulate sources and sound effects out of smaller components with unrelated or ambiguous source-sound relationships, behaviours for a computational model can be designed within a parameter space of perceptual descriptions. Such computational models will be referred to from hereon as *timbre-led models*, as its purpose is to facilitate the external design of behaviours on the basis of an exposed timbre space.

### 3.2.2 Behaviour as Data

Maybe the most obvious way of designing and representing behaviours would be to use a *breakpoint*-based system as found in most digital audio workstations (sometimes referred to as automation curves), or in animation editors where they are referred to as *keyframes*. Figure 3.5 illustrates a behaviour mimicking a 'sword swooshing' action for the wind model introduced above: the centre frequency and gain rise with the movement of the sword (airspeed increases) and the quality factor changes as the sword is tilted (surface area of the air obstruction changes). While these values could have been approximated based on an understanding of physical principles, a similar

result could have been obtained by listening to reference material and adjusting values by trial and error. Making stylistic deviations or embellishments based on listening is easily achieved. For example, the behaviour illustrated in Figure 3.5 would not be physically viable as the gain parameter does not follow the same trajectory as the frequency, but might be deemed more appealing by the designer upon multiple iterations of listening.

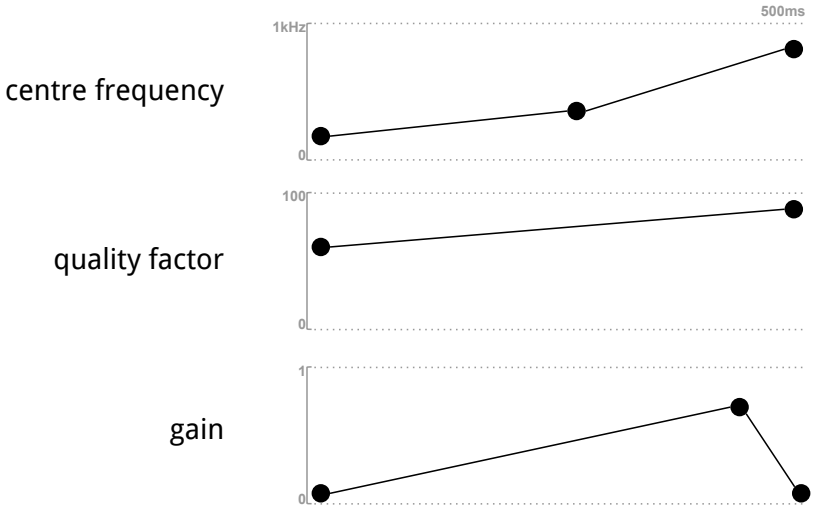


Figure 3.5: Custom 'sword swoosh' behaviour for the perceptual model defined in breakpoints.

The data representing the behaviour has a low informational bandwidth, for example consisting of eight datapoints for half a second in the above example. A much high density of values is unlikely due to the temporal constraints of such trajectories, as will be explored briefly below.

### 3.2.3 Restrictions on the Parameter Space

There are of course limits to the complexity of the parameter space at which it is no longer practically viable to design individual parameter trajectories. A commonly cited example of an exceedingly complex parameter space is in manual implementations of additive synthesis, where dozens of sinusoidal components need to be modulated at the same time in order to emulate sources. Miller famously refers to the 'magical number seven (plus or minus two)' as an approximation of the number of chunks of information that can be held in short-term memory Miller (1956).

### 3.2.4 Temporal Constraints

Fast changes and discontinuities can of course be achieved, but it would counter-productive to consider any movement that occurs as fast as audio rate (for example periodic oscillations) in the navigation of timbre space. Giving rise to new or unexpected psycho-acoustic effects (as in the case of frequency modulation), these sorts of parameter modulations would be more typically integrated with an intentional purpose as part of the signal chain.

### 3.2.5 Design Strategies for Exposing Timbre Space

#### Parameter Exposure by Reduction

As discussed above, the air obstruction model inherently incorporated a perceptual parameter space in the underlying signal chain, expressible as dimensions of 'intensity', 'frequency' and 'resonance'. In this case, a timbre space can be exposed by simply removing additional layers behavioural abstraction that are mapped onto it. This process will be defined here as a process of *parameter exposure by reduction*. An underlying perceptual model already exists due to perceptual considerations in the conception of the model and simply needs to be exposed by means of removing higher-level abstractions.

#### Parameter Exposure by Iteration

Practical synthesis methods, as described above, do not preclude a purely perceptually-driven analysis. As such, it would be possible to derive a 'whistling' or 'swooshing' model based entirely on listening analysis, identifying core perceptual features that are then interpreted in the signal domain.

However, it would be misleading to assume that this is always a straightforward process. What are perceived as independent perceptual dimensions do not always constitute orthogonal spaces when implemented in the signal domain. For example, while interpolation techniques applied in Wessel (1979) are generally applicable to additive synthesis techniques, this is not the case with other signal processes. Yee-King and Roth (2008) applied unsupervised learning techniques to automatically deduce parameter settings for example sounds. While parameter settings were found to converge well on reference examples, the underlying sound model was not capable of facilitating perceptually meaningful interpolation across them. While the signal chain was capable of producing an extremely wide range of sounds, the underlying parameter space was not designed with consideration of perceptually meaningful features that correspond to these sounds, resulting in what Farnell (2011) refers to as disconnected 'islands' in the parameter space. For similar reasons, it would be difficult to deduce a multi-dimensional timbre-space from a single stage of perceptual

analysis when applied to complex sounds.

The next chapter proposes a design strategy for iteratively developing a timbre-space by incorporating the composition of behaviours into the design process. Perceptual features are identified one at a time, undergoing a process of evaluation guided by analytical listening in conjunction with reference sounds, before extending the parameter space further. In other words, the externalisation of behaviour is leveraged to assist the identification of dynamic perceptual features.

## 3.3 Summary

Practical synthesis techniques proposed by Farnell offer a useful alternative to the purely physically-oriented simulation of sound. By incorporating both physical and perceptual analytical strategies into the design process, the designer has the opportunity to limit the model to a subset of behaviour required by the targeted application while also being able to creatively intervene at any stage of the process.

The behavioural components of the sound have been defined here as being fundamentally related to the way that the parameter space of an underlying signal chain is made controllable using a meaningful set of physically viable parameters. The proposed sub-class of *timbre-led models* are conceived on the basis of eliminating all such components from the model in favour of exposing a perceptually meaningful space to control the signal chain. This allows behaviours to be composed as sequences rather than as processes, taking inspiration from graphical animation techniques.

A design process of *parameter exposure by reduction* has been discussed here with reference to the simple air obstruction model, whereby an inherent perceptual space is exposed following the physically informed design process of a model. A further design strategy has been considered, based on a purely perceptually-focused approach to analysis. The next chapter proposes a four-stage iterative design process that incorporates the composition of behavioural sequences as a means of developing a meaningful parametric space. Following this, various control strategies are developed for composing such sequences in real-time using a physical interface. While the successful implementation of real-time performance would speed up the process of composing behaviours, they also have the potential to accommodate expressive nuances in behavioural timbre space. Ultimately, performed sequences could be integrated into an interactive application through dynamic playback and interpolation (see Figure 3.6).

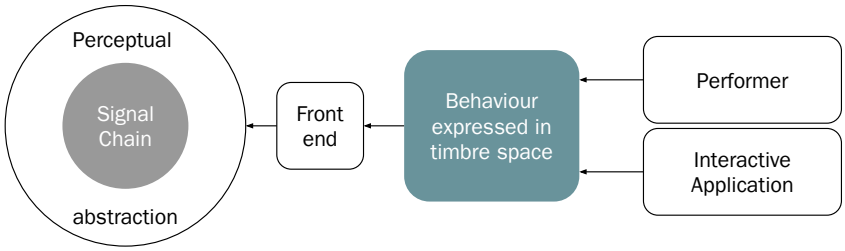


Figure 3.6: Externalisation of behaviour by exposing a timbre space through perceptual abstraction.

# Chapter 4

# Design and Evaluation of a Performable Model of a Creaking Door

## 4.1 Introduction

The last chapter introduced the concept of behaviour in a sound model. While physically realistic behaviour is something that is 'obtained for free' (Farnell, 2011) in a source model, a signal model typically incorporates an explicit representation of the observed behaviour (referred to here as the *behavioural layer*). While there needn't be any clear distinction between behaviour and signal chain in the implementation of numerical models, the proposed class of *timbre-led* models explicitly excludes behavioural properties from the signal chain, relying on a parameter space that describes idiosyncratic perceptual qualities rather than physical variables pertaining to the modelled source. The benefit of this is that behaviours can be designed in a separate process and expressed as data, potentially making it faster to create unique sound effects than if the behaviour was described procedurally within the model.

As discussed in the final section of the previous chapter, one potential way of generating behaviours on-the-fly is to perform them in real-time using physical interfaces. This chapter explores this approach in greater detail, following the development of a timbre-led creaking door model, multiple mapping strategies (or *control layers*) for interacting with its perceptual parameter space and, finally, a user study focused around the rapid generation of sound effects according to evocative narrative scenarios.

Section 4.2 provides a brief overview of stick-slip friction (the physical process underlying the sound associated with creaking doors) and ways of modelling it on

both a source and signal basis. Section 4.3 describes the development process of a timbre-led creaking door model, starting with a basic signal model developed by Farnell (2008) and extending it to incorporate increasingly complex sonic features observed in recorded reference material. Following the design strategy outlined in Chapter 3, each iteration of the development results in an additional parameter (or set of parameters) that can be used to vary the sound output in real-time. Section 4.4 presents three different control layers that transform sensor data from a touch-capacitive surface to three varying parameters of the creaking door model. These include *one-to-one* and *many-to-many* mappings (as commonly applied in HCI and in digital musical instruments) and a novel mapping strategy referred to here as a *physically-inspired control layer*, which emulates the physical behaviour of a bowed string. In addition to this, four metrics for evaluating the performable model (including the aforementioned control layers) are presented: *range*, *nuance*, *repeatability* and *believability*. Finally, Section 4.5 presents the procedure and results of a user study evaluating the performable model involving 15 participants with experience in sound design and musical performance.

## 4.2 Existing Approaches to Modelling Stick-Slip Friction

A creaking door was chosen as a key sound effect to study due to its temporal, timbral and contextual flexibility.

1. The sound is not restricted to any particular temporal morphology (e.g. as opposed to impact or explosion sounds which always consist of a rapid attack and a gradual decay toward silence).

2. The sound can have a lot of timbral variation: squeaky doors can groan, screech, jitter, etc.

3. The causal relationship between sound morphology and physical mechanism (i.e. motion around the hinges) is opaque to the casual listener. This means that when presented with the task of producing the sound for the same physical door two performers are likely to generate different sorts of creaks depending on the context of the sound effect and their own aesthetic preferences.

### 4.2.1 Source Models

The sound of a creaking door is produced by the physical process of stick-slip friction, which refers to an object's periodic sticking and slipping motions as it moves across a surface. Stick-slip friction has been studied in great detail in a variety of fields,

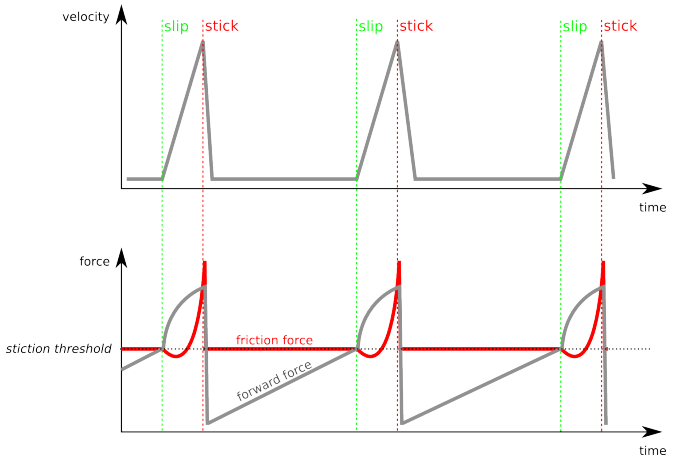


Figure 4.1: Stick-slip friction: dynamic interaction of forces (below) and approximation of resultant velocity (above)

including control theory (Karnopp, 1985), haptics (Sinclair et al., 2011) and seismology (Brace and Byerlee, 1966). A correspondingly large number of models has been developed for different purposes within each field, accounting for salient effects within the particular application or simulation (Berger, 2002). All of these implementations can be categorised into *static* or *dynamic* models of stick-slip friction. Static models only take into account the relative velocity between a moving object and a surface to produce outputs and produce correspondingly limited output in comparison to dynamic models (Serafin, 2004). Dynamic models account for other variable parameters such as perpendicular force and also simulate time-dependent effects.

The periodic sound associated with friction is the result of a dynamic process between two objects with relative motion, illustrated in 4.1. Microscopic surface imperfections cause a resistive force to build up that opposes the vector of a continuous force applied to the moving object. Because friction increases as a function of velocity, the object's force and the friction force eventually cancel each other out and the object's velocity decreases to zero, i.e. 'sticks'. Velocity rapidly increases to a higher value again as the applied force exceeds the friction force (at this point typically referred to as *stiction*) causing the object to 'slip'.

The phenomenon is described in greater detail by Serafin (2004), where a real-time mathematical 'elasto-plastic' model (Dupont et al., 2002) of this dynamic process is implemented, outputting accurate audio samples in response to physical input parameters. Aside from rubbing force and pressure, which can vary in real-time, the model also takes several invariant parameters such as surface roughness, breakaway displacement, and other coefficients specific to the algorithm used to emulate the physical process with a high level of accuracy and detail. Different settings of invari-

ant parameters can cause dramatic changes to the way the physical model reacts to real-time input parameters.

While the feedback resulting from the two variable parameters is natural and intuitive when coupled with an appropriate physical controller, the invariant parameters are less accessible. The models themselves are very abstract; for example, the elasto-plastic friction model is a development of the LuGre model (De Wit et al., 1993), which uses the concept of bristles to simulate microscopic interactions between an uneven surface and a moving object - hence the model exposes seemingly obscure parameters such as 'bristle stiffness' and 'bristle dissipation'. Serafin offers phenomenological descriptions of these parameters, e.g. 'affects the evolution of mode lock-in', 'affects the sound bandwidth' (Serafin, 2004; Delle Monache et al., 2008a), in an attempt to make the control layer more transparent. While, with enough experience and practice, a sound designer can obtain a good understanding of the way parameters affect behaviour and therefore timbral qualities, it remains very difficult to control the sound beyond the intrinsic constraints of the model's physical simulation.

The dynamic variables of *normal force* and *tangential force* lend themselves well to a literal implementation, for example if one were to use an actual door equipped with appropriate sensors to drive the physical parameters of the model. However, it is harder to control specific features in the evolution of the sound, whereby the sound designer is more likely to think in terms of timbral variations such as *pitch* and *roughness*.

### 4.2.2 Non-Dynamical Approaches

As discussed in the previous chapter, it is possible to leverage control over the model's behaviour by replacing the dynamical system with a static but more modular process. Static implementations of dynamical processes are common in digital synthesis, such as the linear time-invariant guitar model by Karjalainen et al. (1998). A more related example can be found in vocalization models. Here, the motion of the vocal folds can be modelled using either parametric or time-variant techniques, where the latter uses lumped-mass models to simulate the dynamic behaviour of the folds as air flows through them (Birkholz et al., 2011) and the former directly synthesises the glottal waveform and/or its derivative (e.g. (Plumpe et al., 1999)) using techniques not too dissimilar from widely-used waveshaping or wavetable synthesis. Similarly, stick-slip friction can be modelled parametrically by synthesising the resulting velocity of the object directly instead of simulating the underlying complex behaviour that has caused it.

The top part of Figure 4.1 illustrates the velocity of an object resulting from the dynamic stick-slip friction phenomenon described above. Interdependencies between the applied force and the friction force between the movable object and a surface result

in regular spikes of velocity, causing the object to be displaced in short impulsive steps. Instead of modelling the force interactions (which would need to take into account physical parameters such as mass of the object, surface viscosity, and so on) the resultant behaviour of velocity over time is modelled independently. The individual spikes can be generated by an impulse train of variable frequency and pulse shape; this can then be passed through a bank of resonant filters to approximate the frequency response of the resonating body. Due to the modularity of this design more features can be added to the model by taking further observations and implementing them into the signal chain.

A basic time-invariant implementation of stick-slip friction, as described above, is presented by Farnell (2008). The regular bursts of velocity resulting from the object's slipping from and sticking against a surface are simulated using an impulse train generator. The output of the impulse generator is then passed through a bank of band pass filters and delay-lines to simulate the effect of the wooden panel. The frequency and amplitude of the impulses relates broadly to the amount of tangential force applied to the object (in this case resulting in the door's rotation around the hinge). Thus an incoming force parameter is scaled to the desired frequency and amplitude ranges. Finally some temporal smoothing is applied to the force parameter to simulate the mass of the door and resulting momentum. These two stages of smoothing and scaling can be thought of as the *behavioural layer* of the door creaking model, while the impulse generator, formant bank and resonators constitute the model's signal processing chain (see Figure 4.2).

Because there is no longer any dynamical interaction between the virtual resonating body and the excitation signal, a lot of the complex behaviour that can be observed in physical models of stick-slip friction is lost. However, this also makes the signal chain more transparent by providing more potential for control over its atomic components. The design process outlined in the following sections serves to illustrate how this can be exploited to extend the model so as to make it more controllable on a perceptually salient level, albeit at the cost of lower or non-existent physical coherence. In other words, the parametrisation of such a model is no longer suitable for directly interfacing with physical input parameters such as velocity and force, but can instead be used to sculpt sound effects in a purely sonic (or *spectromorphological*[1]) context.

### 4.2.3 Towards a Timbre-Led Model

When performing the sound of physical sources as part of a design process one is unlikely to be thinking solely in terms of physical behaviours and might instead think along more perceptually relevant dimensions such as 'brightness', 'pitch', 'harshness'

[1] (Smalley, 1997)

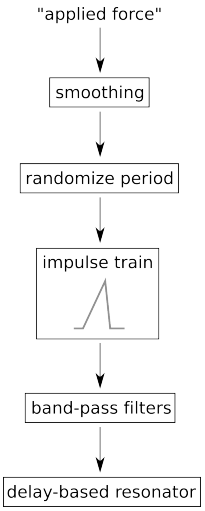


Figure 4.2: Simplified block diagram of Farnell's creaking door model

and 'loudness' (Rocchesso and Fontana, 2003). Dimensions might be relevant to a wide range of people, or they could be highly subjective and describe a particular feature that is important to a given sound designer. Thus a different analytical approach is taken here to extend the model for performance, namely one that takes into account differences in the way something *sounds* rather than the way it *works*.

In one sense Farnell's model can be seen as a special case of a static friction model. Greater amounts of relative angular velocity between the door and its hinge result in shorter intervals between sticking and slipping states, which are observable as higher frequencies. On the other hand it can also be seen as a physically-informed creaking door *sound effect synthesiser*: a combination of atomic signal processing blocks that happens to be very good at generating sounds resembling creaking doors. The model is limited to a single dimension of control, but because of its modular structure it is easier to extend the signal processing chain to account for new timbral features. One can freely switch from a physics-based approach to a sound-design-led approach, implementing new features in the synthesis chain through iterative *phenomenal* analysis (Farnell, 2011).Of course, by following this process the model is more likely to diverge from a physically realistic implementation. Conversely, however, more control can be gained over the sound's timbral morphology independently of corresponding physical processes, while maintaining its sonic identity.

## 4.3 Timbre-Led Model of a Creaking Door

Building on the proposition of *timbre-led* models defined in Chapter 3, a four-stage design process has been implemented. The composition of behaviours in timbre space was incorporated into the design process, matching reference sounds in order to iteratively extend the parameter space of the model.

This process was both conceived and led by the author of this thesis, who has professional experience in sound design for a wide range of media including film, theatre and games.

### 4.3.1 Four-Stage Design Process for a Timbre-Led Model

The following steps were carried out iteratively:

1. Attempt to match a reference sound by programming exposed model parameters over time
2. Identify a perceptual feature that is unaccounted in the existing model
3. Implement and parameterise the missing feature in the signal chain
4. Evaluate the implementation by attempting to match identified feature in reference recording, repeating previous step and this one until an adequate implementation is obtained

Step 1 involves programming the exposed parameters of the latest iteration of the model to match the reference recording as closely as possible. The matching process is complete once the temporal evolution of the feature that the active parameter corresponds to adequately resembles its equivalent behaviour identified in the reference recording. In some instances signal analysis tools might be used to aid this process. For example, spectrum analysis and inspection of the waveform can inform the value and temporal placement of a frequency dependent parameter (e.g. 'pitch'). In other cases (e.g. when the feature is based on a subjective description, and has a correspondingly idiosyncratic implementation in the signal domain), an adequate match comes down to critical listening on behalf of the designer.

Step 2 involves identifying missing features in the sound model based on a listening comparison between the original reference and the programmed output of the sound model. This stage can involve a multitude of analytical approaches, including perceptual, physical or signal analysis (as discussed by Farnell (2011)). Particularly in the case of a highly subjective idiosyncratic feature (e.g. 'screechiness') this may not be well defined in the mind of the designer, in which case multiple and rapid iterations of steps 3 and 4 can lead to a successful implementation.

Step 3 is the implementation of the feature into the signal chain of the model. It requires an understanding of how the observed sound transformation can be translated into the signal domain as an extension to the existing model.

The evaluation process defined in step 4 marks a particular difference to the design strategy proposed by Farnell. On one hand, this may entail improving the implementation based on technical faults observed upon signal analysis (e.g. aliasing artifacts). More pertinent, however, is that the main objective of the evaluation here is the ability of the exposed parameter to accommodate the external definition of temporal behaviours. Therefore, this stage is similar to step 1 (i.e. an attempt to match the reference sound), except that the object of analysis is the technical implementation of the feature in the signal chain.

The key difference to designing specific behavioural abstractions is that the final parameters are representative of a subjective selection of sound qualities rather than an attempt at simulating physical behaviour. Rather than constricting the model to imitate a particular physical process the parameter space is expanded until the model is capable of reproducing all the desired sounds, usually resulting in a larger sound output range than before. An advantage of this approach is that, due to the iterative nature of the parametrisation task, parameters are likely to be linearly independent. In other words, the model designer is unlikely to implement a new parameter that can be recreated using a combination of existing parameters. The drawback, of course, is that the sound output range is more likely to contain unwanted sounds as the model's dimensionality increases due to unforeseen parameter combinations. It should also be noted that the parametrisation becomes highly subjective: what one person holds to be an independent perceptual dimension might not apply to the way another person understands or perceives the sound.

### 4.3.2 Creaking Door Model

A timbre-led model of a creaking door was developed using the four-stage design process outlined in Section 4.3.1.

A set of reference material was compiled, consisting of recordings taken using a portable recorder and contact microphones as well as sounds extracted from professional sample libraries. Five contrasting sequences of continuous creaking were extracted from the recordings and used as references to aid the design process of the sound model. The signal chain was then iteratively extended and parameterised in order to simulate effects heard in the recordings. The model was developed using Puredata but could easily be ported to other DSP environments (e.g. Stowell's Supercollider implementations of Farnell's models (Stowell, 2012)).

A digital audio workstation (DAW) with breakpoint editing functionality was used

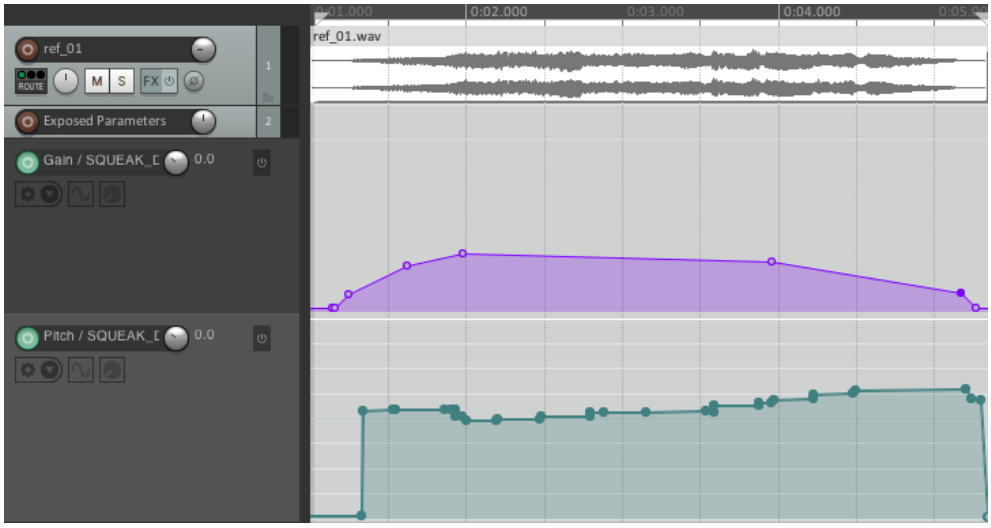


Figure 4.3: A screenshot illustrating the parameter matching workflow in REAPER. Breakpoint envelope output is sent to an instance of the sound model in Puredata, which is developed in parallel.

to compose temporal sequences of variable parameters. In this case REAPER[2] was used to send OpenSoundControl[3] messages to an instance of the model in Puredata (separate lanes of breakpoint curves corresponded to parameters of the model). This made it possible to rapidly match parameter sequences by switching between the reference audio and the sound model output (see Figure 4.3).

The original state of the model was the unaltered implementation described in (Farnell, 2008) (see Figure 4.2 above). This base model consists of a single *applied force* parameter which affects the gain and rate of a pulse generator. The output of the pulse generator is passed through a bank of band-pass filters and delay-based resonators to emulate the resonance of the body. Delay-based resonators follow the *single delay-loop* (SDL) structure (Karjalainen et al., 1998) consisting of a delay line of variable length and a low-pass filter feeding back onto the input signal.

The first iteration of the design process involved the most drastic modifications to the model. This included the exposure of fixed parameters in order to match the source-filter model to the resonant qualities of the recorded doors. *Frequency*, *resonance* and *gain* parameters where exposed for each band-pass filter and delay-line in the corresponding resonator blocks (see Figure 4.4). An additional single set of *feedback* and *damping* values scaled corresponding parameters across all delay-lines in the resonator bank. These parameters were configured once for each reference

[2]http://www.reaper.fm
[3]http://www.cnmat.org/OpenSoundControl

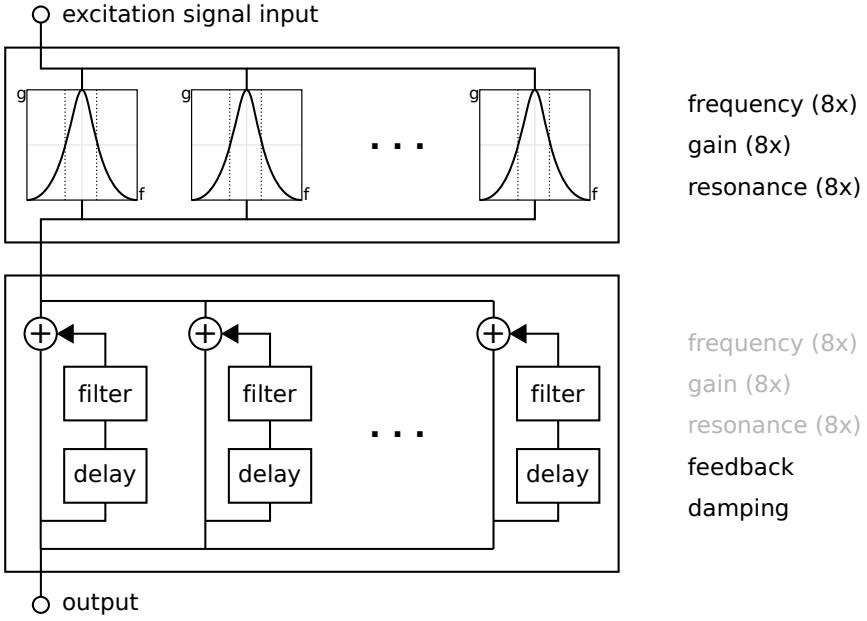


Figure 4.4: Architecture and exposed parameters of the resonator in the squeaky door model, consisting of parallel bandpass filters and single delay-loops.

recording and set by trial and error with the aid of spectrum analysis.

The first varying features that were identified were *pitch* and *gain* of the pulse train source. The original parameter of *applied force* and its corresponding mapping to the signal chain were removed and replaced with independent controls over the rate and gain of the pulse-train generator. Upon a further iteration of the design process the pulse-train generator was replaced with a more elaborate oscillator that provided control over the shape of individual pulses. The pulse width was exposed as a parameter corresponding to the *brightness* of the creaking, which was another identified feature in one of the analysis stages. The dynamic parameters were doubled at one point to account for a second hinge, which was identified in the final two reference tracks consulted in the design process. A second resonator bank was also added in order to implement one of the features identified in two of the reference sounds (*resonator pitch shift*).

Table 4.1 shows a complete list and short description of all the varying parameters exposed in the model, in chronological order of implementation. Table 4.2 shows which reference tracks gave rise to the implementation of these parameters. A simplified block diagram of the final model is illustrated in Figure 4.5.

Because the targeted sound library consisted of a broad range of creaking door sounds the extended model was correspondingly complex - extending the single ex-

| **Parameter** | **Technical Description** |
|---|---|
| Pitch | Controls the pitch of the pulse-train oscillator |
| Gain | Overall gain of the *hinge* output before being passed into the resonator(s) |
| Roughness | Affects gain of low-frequency noise modulating oscillator frequency |
| Brightness | Narrows the pulse width of the pulse-train oscillator |
| Amplitude Modulation Amount | Sets the gain of the amplitude modulation oscillator |
| Amplitude Modulation Rate | Sets the frequency of the amplitude modulation oscillator |
| Noise Gain | Gain of a separate noise generator passed through the resonator(s) |
| Resonator Pitch Shift | Proportionally shifts all the frequencies of the first resonator |
| Second Hinge | An additional *hinge* signal block with parameters 1-7 |

Table 4.1: Variable parameters of creaking door model (in chronological order of exposure)

| **Reference** | **Technical Description** |
|---|---|
| 1 | *Pitch*, *Gain*, *Roughness*, *Brightness*, *AM amount*, *AM rate* |
| 2 | Pitch, Gain, Brightness, *Noise gain* |
| 3 | Pitch, Gain, Roughness, Brightness, *Resonator Pitch Shift* |
| 4 | Pitch (1), Gain (1), Brightness (1), Roughness (1), *Pitch (2)*, *Gain (2)*, *Brightness (2)*, *Roughness (2)*, *AM amount (2)*, *AM rate (2)* |
| 5 | Pitch (1), Gain (1), Brightness (1), Roughness (1), Pitch (2), Gain (2), Resonator Pitch Shift |

Table 4.2: Parameters required to emulate each reference sample in chronological order (italicised text denotes newly exposed parameter in signal chain)

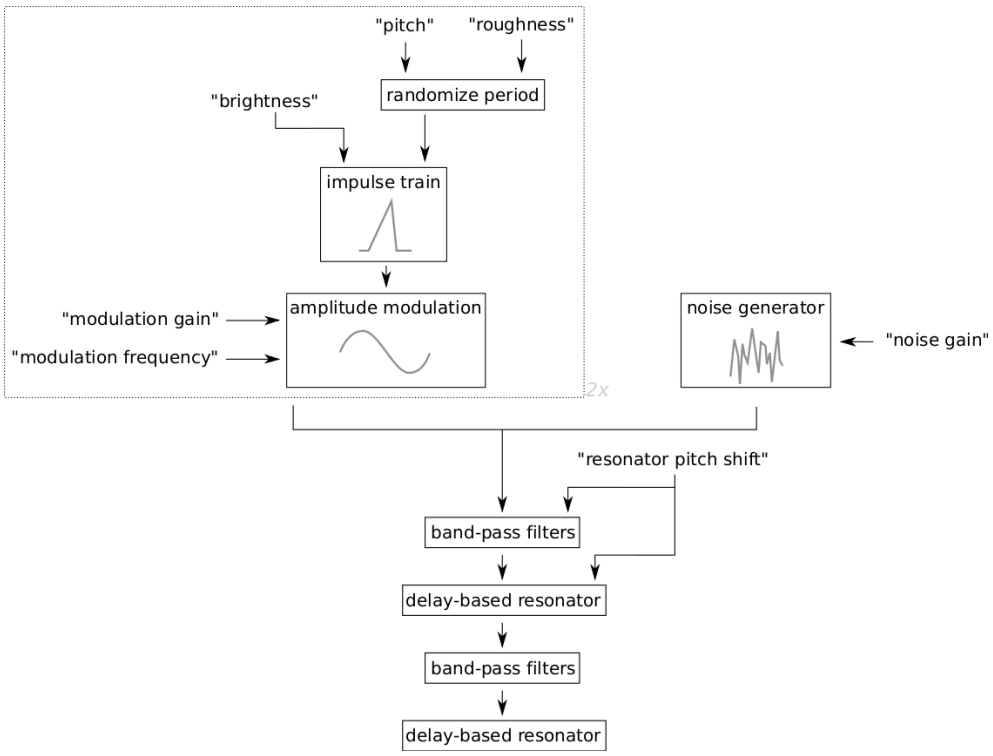


Figure 4.5: Simplified block diagram of the final parametric squeaky door model and variable parameters.

posed parameter of the original model to fourteen variable and fourty-eight fixed parameters. In a typical design scenario a computational model is likely to have a more narrowly defined target sound palette - the library being a set of highly stylised concept sounds rather than arbitrary recordings from the everyday environment - possibly resulting in fewer parameters. However, it should be noted that with further iterations of analysis the corresponding features identified become more exotic as they occur in fewer of the reference sounds. For example, parameters corresponding to *amplitude modulation* and a second hinge were only identified in two of the reference sounds, while *pitch*, *gain* and *brightness* were found to vary throughout the entire set. The original reference sounds and versions emulated through the model at each stage can be found in the on-line supplementary audio-visual materials (see Appendix E.1).

Figure 4.6 shows spectrograms of two stages of the mapping process for the first reference sound. The middle spectrogram visualises the output that was generated following the first iteration of matching analysis, where only two parameters (*pitch* and *gain*) were varying over time. The bottom spectrogram is the final output of the matching process with six varying parameters. A quick visual comparison of these two spectrograms against that of the original reference track reveals the importance of the pitch parameter over other varying parameters.

The next section presents a more detailed overview of each of the exposed parameters. It is worth noting here that while all these parameters were determined based on the perceptually-led design process described above, some of them have been named based on the context of their implementation in the signal chain (e.g. *amplitude modulation* or *gain*).

### 4.3.3 Overview of Retrieved Parameters

#### Pitch

The original model's basic behavioural layer (described in 4.2.2) consists of a single parameter named *velocity*. Its value is mapped linearly to the interval between each pulse (corresponding to a 'slipping' movement). This parameter was renamed to *pitch* and mapped logarithmically to a frequency value that determined the size of the interval. Despite being a fairly trivial modification this eliminates the behavioural layer, replacing it with a parameter that corresponds to the perceived quality of pitch rather than a physical conception of angular velocity and friction. At this stage of the design process the impulse train was also replaced with a band-limited oscillator, to remove frequency aliasing effects observed in the original implementation.

#### Gain

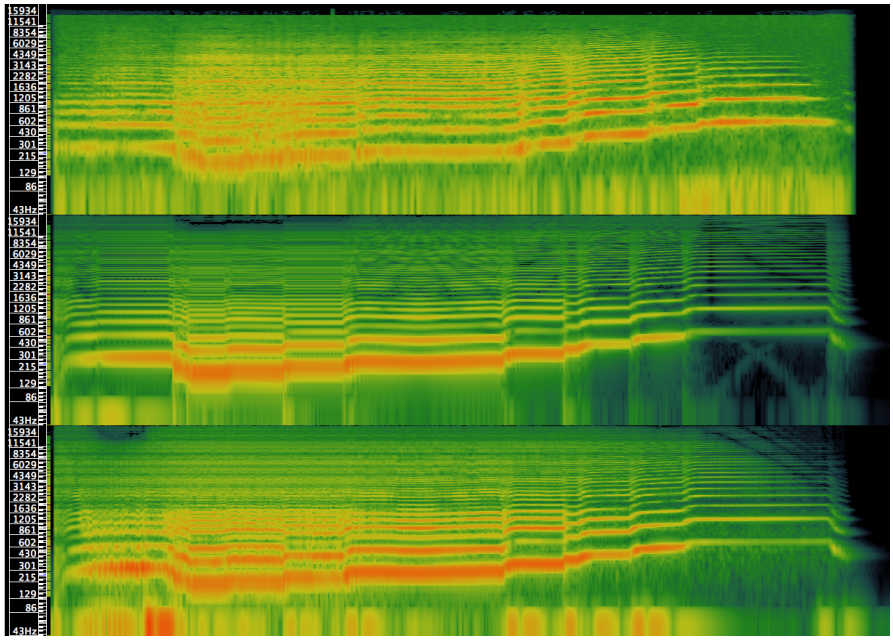


Figure 4.6: Spectrograms of the first reference recording and emulations using model (above: original recording, middle: model output using two varying parameters, bottom: model output using all five varying parameters)

This parameter scales the output level of the impulse train before it is passed into the resonators. The attenuation of high frequencies [4] was retained and made adjustable using a new invariant parameter named *attenuation*. Independent control of gain was necessary in the emulation of almost every door sample consulted, where impulse intensities changed independently of their corresponding intervals.

### Roughness

The original model introduced random variations to intervals between successive impulses, with a range of ten percent of each given interval. The resulting effect is an unstable or *rough* pitch, resembling the sound of screeching tires. The range of randomisation was made adjustable by a parameter named *roughness*, with a linear range of zero to one hundred percent of the current interval time. Some of the door samples contained changes from smooth to much rougher oscillations, which this parameter played an essential part in emulating. The physical counterpart that this relates to is the varying roughness of the surface along which the object moves. In a purely physical model this effect would have to be implemented in the form of a spatial map of surface irregularities (Serafin, 2004).

[4] see Section 4.2.2 in the original implementation described in Farnell (2008)

### Noise Generator Gain

Aside from the periodic sounds emulated by the pulse-train oscillator, the author also observed varying levels of full-spectrum noise in the reference recordings. This was recreated using a white noise generator and its output level is controlled using this top-level parameter. Aside from quasi-periodic sounds door hinges also produce broadband noise due to microscopic surface irregularities. Even most physical implementations rely partly on noise generators to simulate this effect due to the small, random and relatively inconsequential nature of these irregularities (Serafin, 2004; Delle Monache et al., 2008a).

### Brightness

Individual creaks of doors are often distinguished by slight variations in their frequency response while colourations produced by the resonating body remain constant. These can be simulated by shaping the individual impulses produced by the oscillator. The impulse generator was extended making the pulse width (as a percentage of a full cycle) and the length of the attack component (as a percentage of half the pulse width) adjustable. This was later simplified to a single 'brightness' parameter which controlled only the length of the attack component (shorter attack results in a brighter sound). In a physical implementation these effects are highly dependent on material attributes of the moving object and the surface such as viscosity, elasticity and thermoplasticity (Serafin, 2004).

### Multiple Hinges

Many of the door samples contained independently evolving streams of impulses that were most likely caused by the existence of multiple hinges. Each hinge is capable of all of the above-mentioned effects, and thus the amount of controllable parameters increases linearly with the amount of active hinges being emulated. Two pulse-train oscillators, each containing all of the above parameters, were implemented in this model.

### Amplitude Modulation

When one hinge causes strong vibrations in the body of the door, particularly at low frequencies, these can act as an amplitude modulator on the other stream of impulses. This was implemented using separate sine-wave generators oscillating at a variable frequency and amplitude. Though seemingly a cumbersome extension to the model this is a relatively common effect in squeaking doors, particularly when mid-range frequencies (e.g. 200Hz-8kHz) are modulated by a low frequency between 5Hz and 20Hz. This implementation is by no means physically accurate but, as all

the parameters listed here, is capable of convincingly emulating the observed sound effect to the best of the author's judgement.

#### Resonator Pitch-Shift

Another relatively complex sound effect that was observed affects the frequency response of the excitation signal, regardless of variable impulse shapes. Complex spectral patterns are caused by two resonating bodies interacting with each other, more specifically, when the resonant frequencies of one of these bodies are shifted upwards or downwards. This effect can be attributed to the hinge physically deforming as the door moves. Changes in angle cause the door to apply pressure to new parts of the hinge causing a change in tension or a narrowing cavity. In physical terms the effect is highly dependent on the hinge's design and how tightly it is fastened, making this relatively cumbersome to model algorithmically. In this model the effect is easily achieved by shifting all of a resonator's frequencies simultaneously. One drawback is that a second resonator was used to recreate this effect, increasing computation cycles and making it harder to match a given door's frequency response. Some closer psycho-acoustic evaluation might lead to a more simplified emulation of this effect.

### 4.3.4 Summary

In the design of the timbre-led model breakpoint curves were used to compose short sequences of creaking door sounds. This represents a central part of a design strategy that focuses on externalising emulations of physical behaviour from a more perceptually defined signal chain. The result is a perceptual model that is capable of producing a wide range of door creaking sounds by dynamically varying parameters over time. Convincing behaviours can be composed without the need to model potentially complex physical processes. In other words, behaviour is no longer expressed as a self-contained model obtained through careful physical analysis, but as arbitrary sets of data corresponding to morphologies of subjectively meaningful sonic features.

## 4.4 Design and Evaluation of Mapping Strategies for Performing the Creaking Door Model

Having developed a timbre-led model of a creaking door, the focus now shifts from model design to the integration of performance. Chapter 3 proposed the use of physical interfaces to perform parameter changes in real-time (instead of programming them manually on a timeline). On one hand, a successful implementation would drastically reduce the amount of time required to compose sequences. On the other hand, it has the potential to foster more natural exploration and navigation of the

model's parameter space. In the ideal case, performance of timbre-led models would accommodate levels of expressivity observable in non-speech vocalisations and other gestural imitations of sound (as introduced in Section 2.1); enabling meaning to be disclosed through the behavioural properties of the sound without requiring a deep understanding of the processes underlying the source.

Common strategies for controlling non-musical sound follow the *enactive* approach (Essl and O'Modhrain, 2006), whereby parameters extracted from an interface are mapped to sound models in a physically consistent way (see Section 3.1). This is easily achieved with physics-based models in contrast to the timbre-led model described above, which intentionally replaces physical input parameters with a parameter space that is based on perceptual features observed by the designer.

### 4.4.1 Research Questions

As discussed in the previous chapter, the aim of this work is to give the sound designer control over non-physical qualities of the sound in order to compose their own unique behaviours, suiting the given narrative, emotional, musical or other contexts. In the design process described above, idiosyncratic parameters were exposed and programmed over time in order to yield convincing emulations of existing recordings (based on the subjective judgement of the author and with the aid of spectral analysis); thus separating generic acoustic approximations of the sound quality from specific behaviours that are composed externally.

The composition of behaviours was undertaken using a graphical user interface (a breakpoint editor). While this was appropriate in the exercise of matching parameter trajectories to existing recordings, this process is both time-consuming and at odds with the natural performative control sought in this thesis and discussed in the previous chapters.

Assuming that the core signal chain can render an adequate approximation of the acoustic qualities associated with a creaking door, it should be capable of yielding a near infinite variety of potential behaviours by varying all or a subset of the exposed parameters over time. Therefore, would it be possible to compose such behaviours in real-time by mapping streams of physical sensor data to this parameter space? This would allow sound designers to compose believable behaviours using physical movement alone, without requiring a deep understanding of the underlying signal chain.

This forms the basis of the first research question addressed in the following study:

> **RQ1:** ***Is it possible to perform behaviours indistinguishable from manually matched parameters, thereby suggesting a believable result?***

The ability to rapidly perform believable behaviours reduces the need to painstakingly program trajectories using a graphical user interface or other means, therefore providing a significant increase in efficiency. Furthermore, the ability to *intentionally* produce such behaviours to fit a given narrative (or other extra-auditory) context lends the system a degree of expressivity that might otherwise be found in the everyday vocalisation of sounds and the professional work of Foley artists.

There are infinite ways of mapping streams of physical sensor data to a model's parameters and therefore it is crucial to explore how mapping strategies affect the quality of performed behaviours. Four metrics described below in Section 4.4.4 have been chosen to evaluate a set of contrasting mapping strategies (outlined in Section 4.4.3), with the intention of addressing the second research question:

> **RQ2:** ***Which mapping strategies are most conducive to producing believable sounds while offering a high level of control over the parameter space?***

With the aim of shedding light on these questions, three control layers have been developed for interacting with the creaking door model, following the *sensor-mapping-synthesis* paradigm from DMIs (as discussed by Hunt and Wanderley (2003)). A touch-capacitive surface was chosen as a physical sensor, due to its wide familiarity (e.g. from laptops and mobile phones) and high control dimensionality per interaction point (two positional dimensions and contact size). A technical overview of the control layers is provided in Section 4.4.3. Section 4.4.4 describes four evaluation metrics appropriated from the research in control and expressivity of DMIs.

### 4.4.2 Physical Interface

A commercially available touch-capacitive surface (Apple Magic Trackpad) was used as the physical interface in this study. The interface is relatively generic and well-known, providing a lower entry fee to the performer and helping to focus the study on the effects of mapping strategies rather than the physicality of the interface. It also provides a reasonable control bandwidth (three degrees of freedom per finger) while minimising the complexity of the required interaction. No continuous visual feedback is provided while interacting with the physical interface to produce sound.

### 4.4.3 Mapping Strategies

Each of the control layers presented here are based on the same set of input parameters from the touch-capacitive surface (*vertical position, velocity* and *touch size*) and model parameters (*pitch, roughness* and *brightness*). While the sensor can track multiple contact points simultaneously, the physical interaction was constrained to a single finger.

The control layers were developed according to variable levels of complexity and physical consistency (in the sense of the enactive approach described in Section 2.4.4). While a *one-to-one* control layer constitutes the most simple mapping strategy, the *arbitrary-convergent* approach implements a more obfuscated relationship between input and output parameters. The *physically-inspired control layer* is based on a behavioural model of bowed-string interaction, presenting a more physically consistent mapping between physical movements along the surface and perceptual features on the model.

The choice of mapping strategies is to some extent arbitrary as there are an infinite number of possible approaches, including within each chosen sub-category. The goal here has been to choose one emblematic example from three well-known categories in the fields of musical and sonic interaction design. A more detailed description of each mapping strategy is provided below. Precise details of each implementation have been tuned subjectively by the author in order to provide a plausible mapping within each category. Technical implementations of each mapping can be found in the supplementary materials.

A video demonstrating each of these control layers can be found in the on-line supplementary audio-visual materials (see Appendix E.2).

#### One-to-One Mapping

This control layer implements the most simple mapping strategy, wherein each input parameter has independent control over a synthesis parameter (see Figure 4.7(a)). Vertical position controls pitch, where higher pitches are produced when placing a finger at the top of the touch surface. Touch size is mapped linearly to roughness, where a larger value results in a higher level of roughness. Finally, velocity controls brightness (the attack component of the impulse train), where a faster velocity results in a duller sound.

#### Physically Inspired Control Layer

Drawing from enactive interaction strategies outlined in Chapter 2 a mapping based on a physical control metaphor was designed, which is referred to here as a *Physically Inspired Control Layer* (PhICL). The name takes inspiration from the set of behavioural sound models developed by Cook (1997) under the umbrella term of *Physically-Informed Sonic Modelling (PhISM)*. The motivation behind implementing this here was to include a more physically-viable interaction mechanism among the set of control layers explored.

The control layer is loosely based on the interaction between a bow and a string and can be seen as an independent behavioural component placed between the physical control input and the sound model. As mentioned above, the physical mechanism

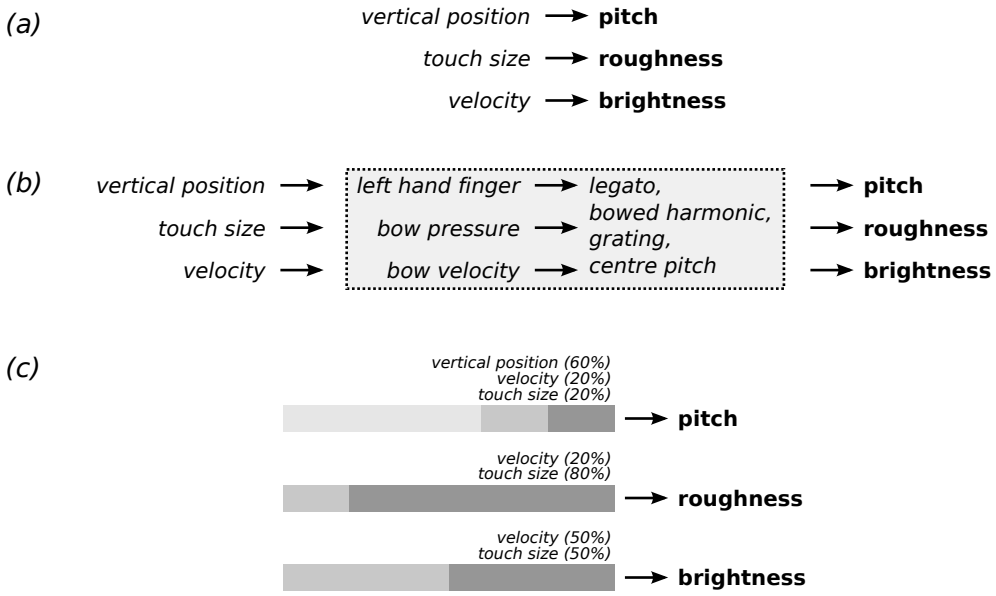


Figure 4.7: Three mapping strategies applied to the control of the creaking door: (a) *one-to-one*, (b) *PhICL*, (c) *convergent*

behind timbral variations of creaking doors is relatively opaque and multi-faceted.On the other hand, a bowed string can produce sounds of similar complexity through comparatively transparent relationships between input and output parameters. On behalf of the sound model, some of these effects can be modelled using similar perceptual descriptions of 'pitch', 'brightness' and 'roughness'. In terms of interaction, the actions of moving a bow across a string with varying degrees of pressure can be assimilated broadly to the actions of rubbing a finger across a surface and varying the contact area.

According to Schelleng (1973), the sound of a bowed string can vary between three states of 'raucous', 'normal' and 'higher modes' depending on how amounts of bow velocity and pressure are combined. The 'raucous' state refers to aperiodic, low-frequency oscillations of the string caused most easily by bowing the string at low velocities and high amounts of pressure. The string is in a 'normal' state when optimal amounts of pressure and velocity cause it to oscillate at its fundamental frequency. The state of 'higher modes' refers to higher modes of oscillation that are activated when too little pressure is applied to the bow while moving it at normal or high velocities, resulting in a higher-pitched and warmer sound. The fundamental mode of oscillation is controlled by varying the length of the resonating segment of the string (e.g. moving the left hand along the neck of a violin). The resulting sounds can be approximated using the three chosen synthesis parameters of *pitch*, *brightness*

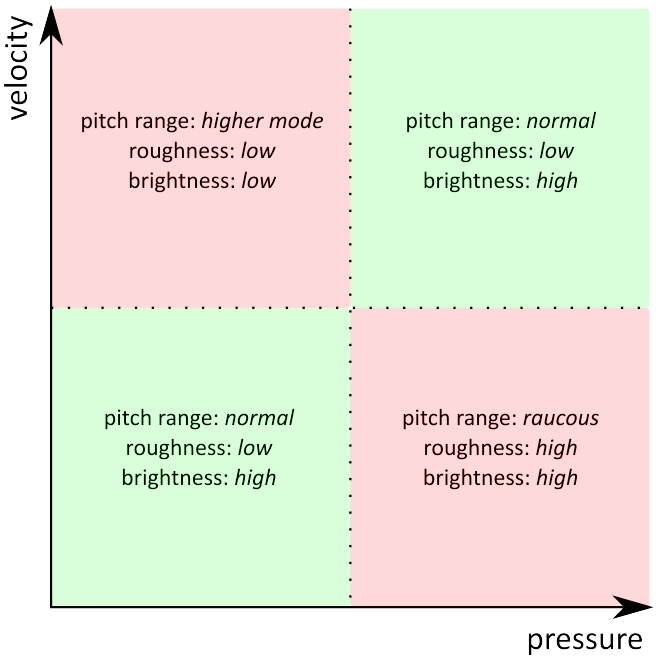


Figure 4.8: Physically Inspired Control Layer: Relationship between virtual bow pressure and velocity and perceptual parameters of the door creaking model

and *roughness*. As illustrated in Figure 4.8, given a fundamental vibration frequency, steady oscillation associated with the 'normal' state can be achieved by applying the right combination of pressure and velocity.

These relationships were implemented as a simplified dynamical behavioural model illustrated in Figure 4.9. Relationships between contact velocity and each perceptual parameter are divided into three Segments (labelled *A, B* and *C*), which are equal at medium amounts of contact area. As the contact area increases, Segment A increases in size, meaning that higher velocity values are required in order reach Segments B and C. At lower levels of contact area the opposite expansions and contractions occur, making it easer to 'overshoot' the fundamental frequency. Absolute vertical position of the contact point is mapped to the fundamental pitch of the model, maintaining the orientation from the previous one-to-one mapping (i.e. up for a higher pitch). The pitch is doubled by an octave when in the 'higher mode' state.

Hysteresis at the segment thresholds was introduced in order to maintain controllability over the model, which would otherwise cause noticeable sporadic switches in states. Effects of hysteresis (or *hysteresis cycles*) have been observed in mechanical studies of bow-string interaction and are commonly associated with dynamic thermal properties of resin applied to the bow (McIntyre and Woodhouse, 1986). These effects are observable in *elasto-plastic* models of stick-slip friction (Serafin et al., 2003). The

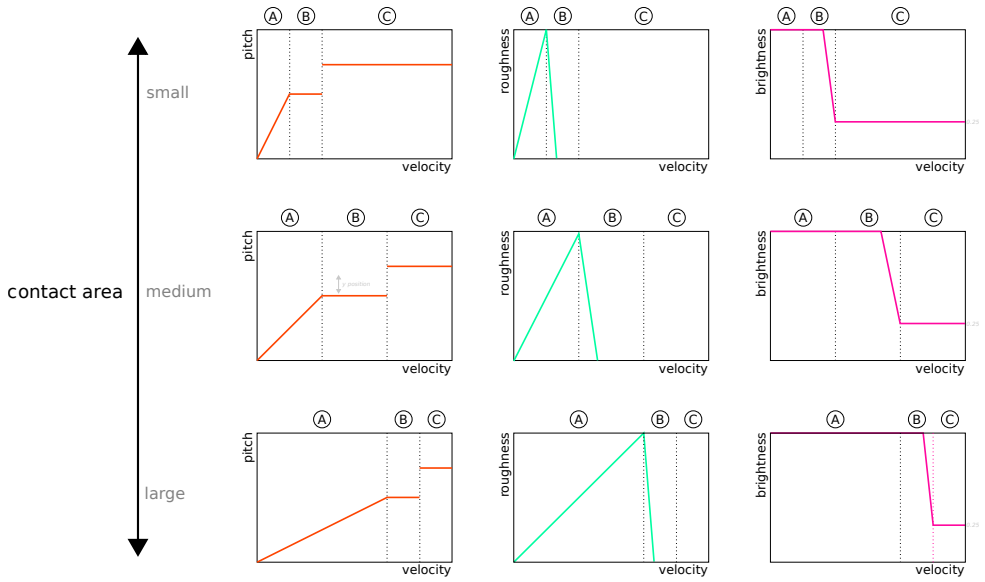


Figure 4.9: Dynamic switching between racous (A), normal (B) and higher mode (C) states in the bowed-string PhICL

simplified implementation of this effect is illustrated in Figure 4.10. Upon crossing from Segment *A* to Segment *B*, Threshold *T1* is shifted back, requiring a greater fall in velocity in order to transition back to Segment *A*; the same applies to Segments *B, C* and Threshold *T2*.

### Arbitrary Convergent Mapping

If a one-to-one mapping strategy is the most simple way of mapping control to synthesis parameters, a convergent mapping approach lies at the opposite pole of mapping complexities suggested by Rovan et al in (Rovan et al., 1997), now more commonly

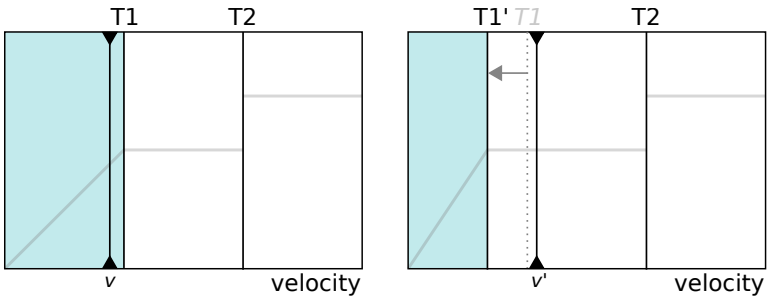


Figure 4.10: Hysteresis at state transitions in the bowed-string PhICL.

| | Vertical pos. | Velocity | Touch size |
|---|---|---|---|
| Pitch | 0.6 | 0.2 | 0.2 |
| Roughness | 0 | 0.2 | 0.8 |
| Brightness | 0 | 0.5 | 0.5 |

Table 4.3: Weightings of input parameters in convergent mapping strategy

referred to as many-to-one or many-to-many mappings (Hunt and Wanderley, 2003). Here each synthesis parameter is controlled by more than one single control dimension. While the PhICL can also be understood as a convergent mapping, this layer is a more arbitrary and simple implementation in that control parameters are simply added together (after being normalized and individually weighted) to produce an output parameter. To ensure a minimum level of playability each input parameter was assigned a different weighting based roughly on the relationship between parameters in the PhICL (see Table 4.3).

### 4.4.4 Four Metrics for the Performance of Environmental Sound

To the best of the author's knowledge, there are currently no established means to evaluate how well an interface lends itself to the expressive human performance of a given environmental sound. Borrowing from the field of digital musical instrument design, four metrics are proposed for the evaluation of the performable timbre-led model.

#### Nuance and Range

Jordà (2005) describes three levels of 'diversity', which are presented as a new take on recurring concepts within the field of digital musical instrument design previously referred to as the 'versatility' or 'flexibility' of an instrument. The first category, *macro-diversity*, refers to the ability of an instrument to perform in different contexts or styles. For example, a guitar has a higher level of macro-diversity than an oboe because it can be applied within a wide range of musical styles and public contexts. The next category is *mid-diversity*, which is most succinctly described as how distinguishable two different pieces of music played on the same instrument can be. An instrument with very low mid-diversity will sound as though it is always playing the same piece of music. Finally, *micro-diversity* is the instrument's ability to control subtle nuances of a musical sound - in other words, how different two performances of the same piece can be.

While these three metrics were specifically conceived to describe musical instruments, they can be recontextualised to apply to the performance of timbre-led models. Micro-diversity then refers to the ability to control subtle nuances of a performed sonic

behaviour. It would, however, be misleading to base this metric on quantifiable descriptions of performances based solely on the model's parameter space. The designer is likely to be performing sounds based on behavioural descriptions that transcend the perceptual parameter space of the underlying model. In other words, this metric should not only describe how well a performer can vary a sound according to dimensions of 'brightness', 'pitch' and 'roughness' but, for example, whether they are capable of performing multiple variations of what they would classify as a 'groaning door'. Mid-diversity, on the other hand, refers to the behavioural breadth that the performable model can produce: can the sound produce a multitude of behaviours associated with a door, or only sounds that are associated with 'groaning'? The likeliest interpretation of macro-diversity in this context would be the amount of different sources that can be represented by the performable model (much in the same way that a Foley artist might use a 'prop' to emulate a multitude of unrelated sources).

For sake of clarity, micro-diversity and mid-diversity will be referred to from hereon as *nuance* and *range* in the context of performable models. The above interpretation of macro-diversity would be an interesting aspect to study but falls outwith the scope of this study, which has been constrained to the performance of sounds associable with creaking doors. A more suitable term for this might be *source appropriability*. Outcomes of the soundtrack synchronisation study presented in Chapter 6 include instances where performable models were appropriated to emulate unexpected sources.

#### Repeatability

Controllability has long been regarded as an important metric in the evaluation of musical interfaces (Wanderley and Orio, 2002). It has been inspired by usability metrics in HCI, and is equally applicable here. In the evaluation of DMIs, it refers to the ability to maintain control over the timing and features of a musical trajectory and is equatable to *repeatability* (the accuracy at which a given sound can be repeated using the same interface), which is the preferred term that will be used here. This metric can easily be evaluated by testing the accuracy at which a previously performed sound can be repeated on the same interface.

#### Believability

As mentioned above, a negative byproduct of the iterative design process for timbre-led models is that the resulting high-dimensional space is likely to contain parameter configurations that bear no relation to the modelled source. This of course does not only apply to static configurations of parameters but also to their temporal behaviour and potential interdependencies. Defining a space of parameter settings (and movements) that always results in the believable emulation of a source's behaviour

is probably impossible to achieve in a high-dimensional model like the creaking door. The ability to produce believable behaviours can of course be guaranteed through the process of successfully matching parameter trajectories to reference sounds, but it is unclear how well these discrete examples of behaviour can be extrapolated to a larger space. For these reasons a *believability* metric should ideally be incorporated into the evaluation of a performable model.

Due to the perceptual nature of the timbre-led model believability cannot be equated to realism, which is why it is not testable by purely physical or mathematical means (e.g. similarity to the output of a physical model). Arguably, even if a quantifiable metric of physical realism was sought, it would only be constrained by the scientific models underlying the source that is being referenced. Furthermore, this would only apply to models that are based on real reference sources. Therefore believability should ideally be tested as part of the listener's experience *within its intended context or environment.*

Finally, it should be reiterated here that believability does not equate to realism or acoustic fidelity. The above four-stage design process was conceived with the intention to separate acoustic approximations of the source from its temporal behaviours - where the former is integrated deeply into the signal chain and the latter is expressed within the perceptual parameter space defined by the model designer. Here, then, believability refers specifically to the behavioural qualities of the performed sounds, and whether a performance is equally good at conveying a particular behaviour as a manually programmed set of parameter changes.

This rests on the assumption that the signal model does indeed produce an adequate approximation of a creaking door. Within this work, this assumption is backed up by spectral consistency in the analytical process, the author's personal sound design experience and a design procedure built upon pioneering scholarship in the field.

## 4.5 Evaluation Study

### 4.5.1 Objectives

The research questions that this study seeks to answer have already been introduced at the beginning of Section 4.4.3. They will be briefly restated here:

1. Is it possible to perform behaviours indistinguishable from manually matched parameters, thereby suggesting a believable result?

2. Which mapping strategies are most conducive to producing believable sounds while offering a high level of control over the parameter space?

An experimental procedure was developed to study the effects of each of the three

control strategies on the performance of creaking door sound effects in a narrative context. The purpose of the procedure was to collect a mixture of quantitative and qualitative data from which an approximation of the metrics described above could be deduced. Results would shed light on some of the strengths and weaknesses across the set of mapping approaches applied.

Perhaps the most important of the two research questions is the first one. On one hand it serves as a means of evaluating the legitimacy of timbre-led models and the proposed four-stage design process. It would point out whether the known behavioural range of the model (expressed as discrete successful emulations of reference sounds) can be extrapolated to accommodate a wide range of believable behaviours without the requirement of audible reference material. Moreover, a positive answer to this research question would demonstrate that it is possible to design such behaviours on the fly, by leveraging human performance.

#### Participants

18 people (7 female, 11 male) aged between 27 and 40 participated in the study. 12 participants had experience in sound design ('recording, editing, implementing and/or producing sound effects for film, stage, games, radio, etc.'). The remaining participants had experience editing and recording sound or played a musical instrument for longer than five years.

### 4.5.2 Environment and Setup

The study consisted of two parts. The first part was a performance study in which each participant was asked to perform a variety of creaking door sounds according to a pre-defined set of tasks. This set of tasks was repeated for nine different narrative scenarios devised for the study. The second part was a listening study in which doors performed by the previous participant had to be distinguished from sequences that were matched to recordings of real doors (corresponding to narrative scenarios from the previous task).

The experiment was carried out in a controlled, quiet environment and participants wore headphones throughout. Tasks were given to the user through an automated graphical environment that was navigated using a specially labelled keyboard, allowing each participant to carry out the experiment without the need for supervision. Nonetheless, participants were given the opportunity to ask questions at any point throughout the study in the case that instructions were unclear. In the first part of the study sounds were performed on the touch sensor which was attached to the table in front of the keyboard. Participants were instructed to only use one finger to generate sounds (to avoid attempts at using common 'swipe' or 'pinch' gestures).

### 4.5.3 Data Collection

Control and synthesis parameters were logged throughout the first part of the study.

Additional qualitative data was collected in a survey consisting of a mixture of Likert-style ratings, binary questions and opportunities to elaborate in comments.

### 4.5.4 Procedure

#### Part 1

In the first part of the experiment participants were sequentially presented with nine imaginary scenarios for which they were asked to perform the sound of a squeaky or creaking door. Each scenario consists of 2-4 sentences written in evocative language to encourage maximum variation for each performed sound effect, for example:

> *Two shifty characters are closing a deal in a secret room in the basement of a bar. As one of them hands a briefcase over the door unexpectedly opens. Both characters instantly pull guns out of their pockets and aim at the door, looking at each other suspiciously. Perform the sound of the door.*

Invariant parameters of the sound model as well as room impulse responses were loaded as pre-defined presets for each scenario corresponding to the type of door and space specified by the scenario. Before being presented with the first scenario participants were given a tutorial on the current control layer (referred to in the study as an 'interface'). The tutorial consisted of no more than four simple images accompanied by text demonstrating how finger movements affect the sound quality (e.g. "Large touch area and low movement speed results in a low-pitched unstable squeak", see Figure 4.11). Participants were able to interact with the touch sensor while navigating through the tutorial. They were then shown the first task scenario and given the opportunity to explore and practise sounds before being asked to record the targeted sound effect three times. Participants were then asked to choose their favourite performance (given the opportunity to listen back to each one). Next, they were asked to repeat the same sound five times, attempting to match the original one as accurately as possible. Following this they were asked to repeat the sound again at a higher or lower pitch, or a faster or slower speed (chosen randomly for each scenario).

The procedure was repeated for each narrative scenario. Control layers were switched after the third and sixth tasks and participants were given the corresponding tutorial before proceeding to the following tasks.

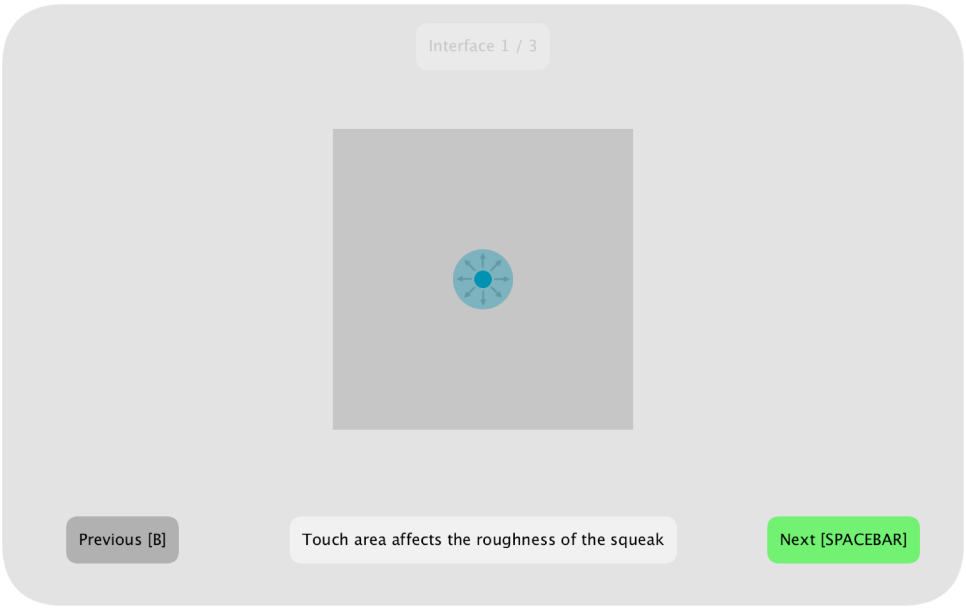


Figure 4.11: A screenshot from the study's graphical user interface describing an interactional dimension of the *one-to-one* control layer

### Part 2

In the second part of the experiment participants were instructed to listen to a radio play that was specially written and produced for this study. The script is presented in Appendix A. The radio play was fourteen minutes long and centred around two characters viewing an old house that used to belong to one of the character's relatives. All dialogue and additional Foley tracks were treated with convolution reverb corresponding to each room that is visited by the characters. The play contained nine occurrences of creaking doors, windows and gates, which physically correspond to the ones featured in the first part of the study (e.g. 'iron gate', 'stiff window', 'light wooden door', etc.). While invariant parameter settings were the same as in their corresponding performance task scenarios, gain and reverb settings were adjusted to blend in naturally with the sound image of the radio play. This was carried out by the author, who also recorded and produced the radio play.

For each participant six out of nine randomly selected door sounds were taken from the favourite performance of the previous participant for the corresponding door type. The remaining three were modelled on recordings of real doors (extracted from sample libraries and own recordings). The recordings were matched by the author using a breakpoint editor as described in Section 4.3. The doors were matched using no other variant parameters than the three that were controllable with the physical controller.

The participant was explained the difference between a door sounds that were performed and those that have been modelled on a pre-existing recording before listening to the radio play. After each time that a door sound occurred in the radio play the participant was asked to choose whether the sound effect was performed or based on a real recording by pressing a corresponding key. This decision was encouraged to be made while the play continued (pausing after five seconds of inactivity) - a flashing screen prompted a quick response by the participant.

### 4.5.5 Survey

A survey comprising of three sections supplemented the practical phases of the study. A further section was completed at the end of the study, after completing the listening task.

#### Sections 1-2

The first section collected general information (such as age, gender and experience) The second section included sets of 7-point Likert-scale prompts that were answered after the tasks for each control layer were completed:

- 'Overall the interface was capable of producing a wide breadth of door sounds'
- 'The interface gave me a high degree of control over subtle nuances of the sound (e.g. while refining the sound for a given scenario)'
- 'I found it easy to repeat the sounds during the repetition task'
- 'Overall I was satisfied with the sounds I created for each scenario using this interface'

These prompts corresponded respectively to the three performance metrics of *range*, *nuance* and *repeatability*, and their overall satisfaction with the sounds they produced on the given interface.

#### Section 3

This section was filled in after all performance tasks were completed. This consisted of six questions asking participants to identify which of the three interfaces:

- was their favourite
- produced the best sounds for the given tasks
- was most enjoyable to play

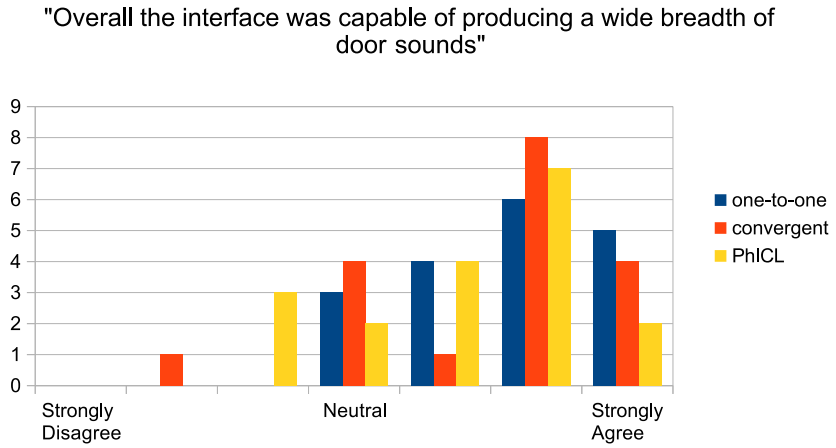


Figure 4.12: Likert-scale responses for *range*.

- was considered the most challenging
- was considered the least challenging

Participants were encouraged to elaborate on their choices in free-text responses.

**Section 4**

The final section of the survey comprised of three further 7-point Likert-scale prompts that were completed at the end of the listening task:

- 'Overall I was confident in my judgement of whether the sound effects were performed or based on a real recording'
- 'The door squeaks that I thought to be based on real recordings were believable'
- 'The door squeaks that I thought to be originally performed by a human were believable'

## 4.6 Results

### 4.6.1 Range

Figure 4.12 shows the Likert-scale responses to the prompt "Overall the interface was capable of producing a wide breadth of door sounds". Each of the three control layers received similarly high ratings. A Kruskal-Wallis test showed no significant difference according to the type of mapping ($H_{adj} = 1.725$, $p = 0.422$).

| | Mean Distance | S.D. |
|---|---|---|
| One-to-one | 1.222 | 0.687 |
| Many-to-many | 0.965 | 0.641 |
| PhICL | 0.282 | 0.255 |

Table 4.4: Distance metrics associated with sequence diversity across narrative contexts. Higher numbers correspond to greater measures of diversity.

This is interesting when viewed in the light of the actual dimensional space that is made accessible by each control layer. Figure 4.13 contains three scatter plots of all the performed sequences (discounting repeats) for each mapping strategy. While this may not be the most precise representation of the parameter space navigable by each control layer, it suffices to make some salient observations. The one-to-one mapping has the most evenly distributed points, which reflects the simple nature of the mapping: any combination of values for the three parameters is reachable by varying each of the three control dimensions. The parameter space navigable using the PhICL, on the other hand, is clearly constrained: it is not possible to combine arbitrary values of *brightness* and *roughness* as an inherent property of the underlying behavioural model (see Section 4.4.3). The convergent control layer is constrained to a much lower pitch range, though this is not an inherent limitation of the mapping strategy. It would be possible to access higher values of pitch by making fast movements along the upper edge of the touch sensor with a high contact area, however this plot suggests that participants avoided doing this.

As noted in its definition in 4.4.4, *range* is not a measure of the parameter space itself but of how much performed sounds tend to vary in different contexts. One approach to measuring this would be to obtain a measure of how much the performed trajectories vary across different tasks for each interface. In an attempt to test this Gaussian Mixed Models (GMM) with ten weighted components were fitted to the parameter data of each task. For each task the log probability of each data point was then calculated under the GMM model of every other task. Finally, normalized mean log probabilities were calculated for each interface and used to generate distance metrics (see Table 4.4). Results correspond roughly to the observations made in the scatter-plots above, with the one-to-one mapping scoring the highest.

### 4.6.2 Repeatability

The convergent control layer received the more positive responses to the prompt “I found it easy to repeat the sounds during the repetition task", while responses for the remaining two were more varied (see Figure 4.14). A Kruskal-Wallace test found no significance within the overall responses ($H_{adj} = 3.529$, $p = 0.171$), however paired T-tests across responses to each interface reveal higher homogeneity between the first

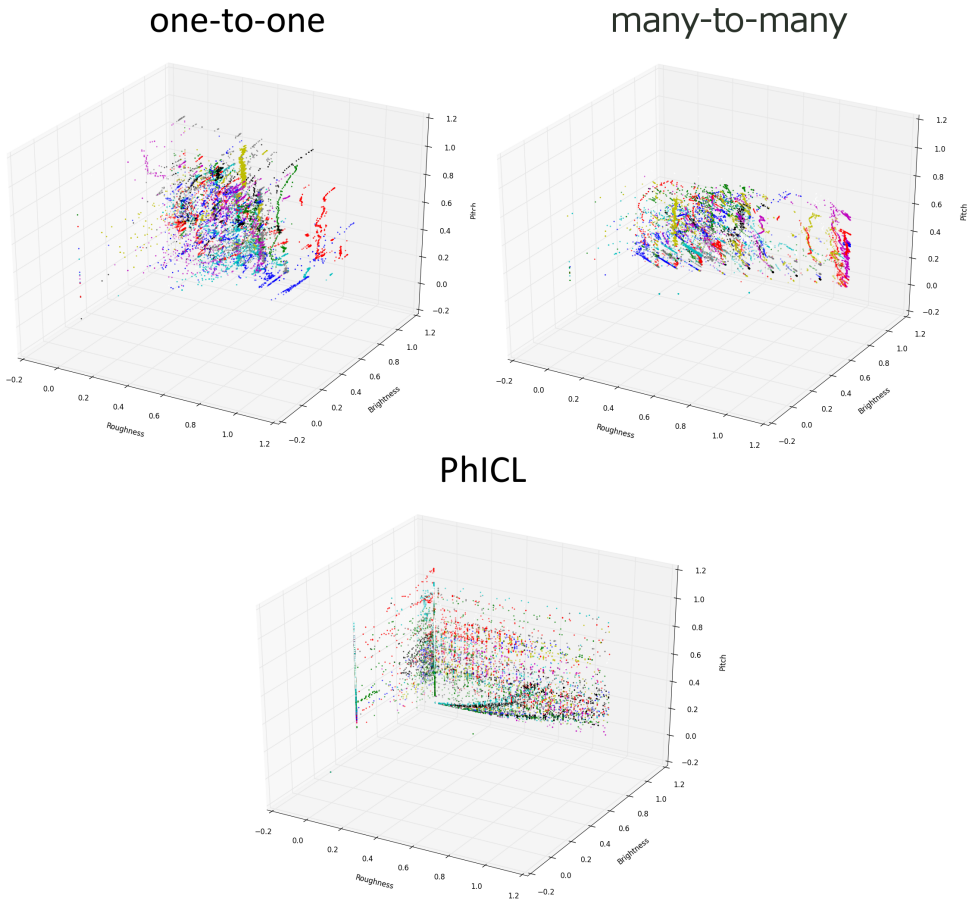


Figure 4.13: Scatter plots of all performance data for each interface. Colours correspond to different tasks.

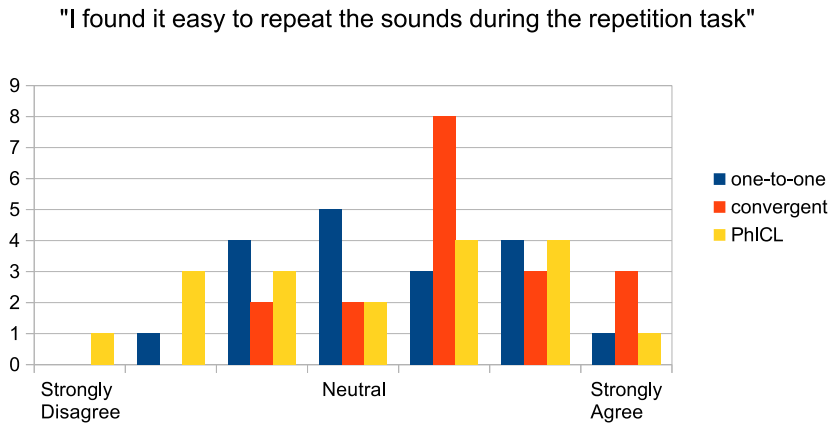


Figure 4.14: Likert-scale responses for *repeatability*.

| | Rel. rating | Mean Alignm. Cost | S.D. |
|---|---|---|---|
| One-to-one | 2.294 | 3.858 | 9.720 |
| Many-to-many | 1.000 | 1.682 | 4.726 |
| PhICL | 3.340 | 5.719 | 8.122 |

Table 4.5: Results from repeatability test

(one-to-one) and third (PhICL) interfaces ($p = 0.601$) than between the first and second (convergent) ($p = 0.103$), and the second and third ($p = 0.059$).

The relative success of repetition across interfaces was measured in the logged sensor data by applying Dynamic Time Warping (DTW) between repeated sequences and their corresponding original performances. This process was carried out for repetition tasks across participants and a mean minimum alignment cost was calculated. The results (see Table 4.5) suggest greater success rates for repeatability in the convergent layer.

### 4.6.3 Nuance

Sensor data relevant to this metric was collected in the extended repetition task in which participants were asked to repeat a sound that they previously performed at a higher or lower pitch or at a faster or slower speed. By discounting pitch and time differences and measuring the divergence between original and repeated tasks a measure can be obtained of how easily a person can vary one aspect of sound without affecting the rest of the parameter space. Larger divergences would suggest a lower success rate in keeping all parameters constant while attempting to vary one.

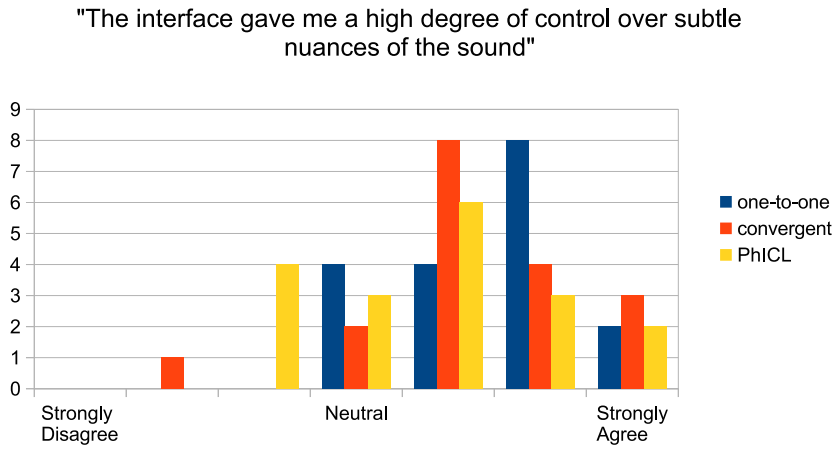


Figure 4.15: Likert-scale responses for *nuance*.

| | Rel. rating | Mean Alignm. Cost | S.D. |
|---|---|---|---|
| One-to-one | 1.583 | 4.6025 | 8.1401 |
| Many-to-many | 1.000 | 2.9082 | 9.5846 |
| PhICL | 2.799 | 8.1401 | 7.5321 |

Table 4.6: Results from test for *nuance*. Lower number corresponds to a higher success rate.

DTW was applied again to measure similarity between original and repeated synthesis parameters. Pitch parameters in repeated sequences were vertically aligned to the corresponding originals in order to discount intentional differences resulting from the pitch-variation tasks. Results show lower alignment costs for the convergent control layer and a greater divergence for the PhICL (see Table 4.6).

As discussed in Section 4.4.4, constraining an evaluation of this metric to the model's parameter space does not necessarily account for all of the properties that cause the performer to perceive nuanced control over the sound. This is because the performer is likely to be thinking about behavioural concepts that transcend the parameter space of the model, considering varying levels of 'screechiness', for example, instead of a specific means of varying the pitch. Figure 4.15 shows participant responses to the prompt "The interface gave me a high degree of control over subtle nuances of the sound". All control layers received mostly positive responses with no significant differences found across each set ($H_{adj} = 2.929$, $p = 0.231$).

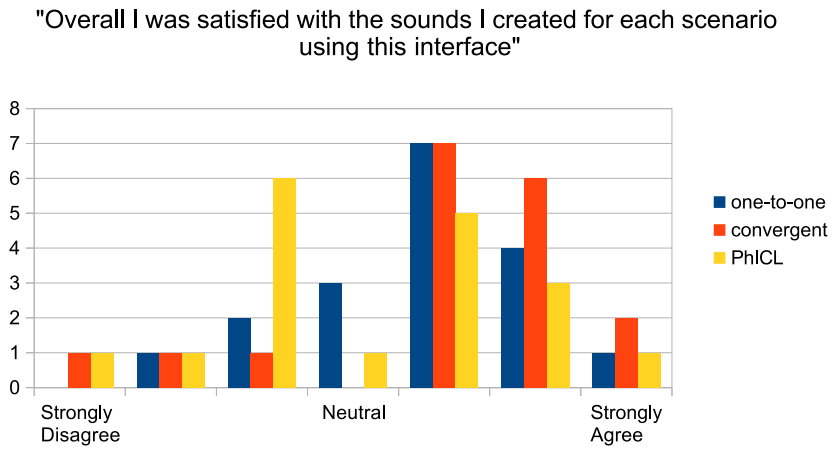


Figure 4.16: Likert-scale responses for satisfaction with performed sounds.

| | one-to-one | convergent | PhICL |
|---|---|---|---|
| was favourite | 7 | **9** | 2 |
| produced best sounds | 6 | **7** | 3 |
| most enjoyable | **8** | 7 | 3 |
| most challenging | 4 | 4 | **8** |
| least challenging | 5 | **7** | 4 |

Table 4.7: High-level survey responses comparing the control layers.

### 4.6.4 General Preferences over Control Layers

Responses to the prompt "Overall I was satisfied with the sounds I created for each scenario using this interface" show high ratings for the one-to-one and convergent control layers and larger number of negative ratings for the PhICL (see Figure 4.16.

Some further observations can be made in responses to more general prompts about the control layers (see Table 4.7. The convergent control layer was found to be the preferred and least challenging one among participants, and was also considered to produce the best sounds for the given scenarios. Some participants commented on the use of contact area to control 'additional characteristics of the sound' with one participant claiming that this lead the interface to be more 'expressive' and 'intuitive'.

The plurality of participants considered the *PhICL* to be the most challenging control strategy. One participant referred to this as an 'expert' interface among other participants who claimed they would get better results with more practice.

The one-to-one mapping received the highest ratings for enjoyability. One participant claimed that 'this seemed like the most intuitive' control layer, while others

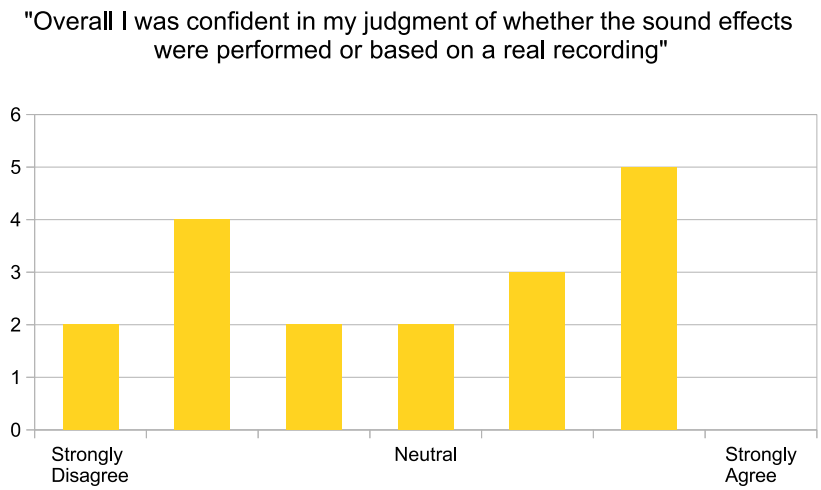


Figure 4.17: Participant ratings of confidence in their ability to distinguish performed sounds in the listening study.

referred specifically to the ease of controlling pitch. One participant attributed this to the fact that no movement was required to produce a steady pitch.

### 4.6.5 Believability

The believability of the sounds produced by the performable model was examined in a listening context. At the point of carrying out the listening study participants had familiarised themselves with the sound model and the three control layers through by completing first part of the study. Thus, any consistent traits of the sound that were attributed to the constraints of a particular interface would likely have resulted in more positive identifications of performed sequences during the listening study. Table 4.8 shows the percentage of correct identifications for each sequence type.

Significance tests were calculated according to binomial distribution.Listeners could correctly identify overall performed sounds with better than chance accuracy ($p = 0.022$), driven mainly by the convergent layer, where listeners were able to distinguish sounds from those based on recordings ($p = 0.012$). The other two mappings showed no significant difference ($p = 0.189$, $p = 0.229$). There was also no significant ability to correctly identify sounds that were based on recordings ($p = 0.276$).

Figure 4.17 shows responses to the prompt "Overall I was confident in my judgment of whether the sound effects were performed or based on a real recording", which were widely distributed.

One participant claimed that if they 'hadn't done the previous sections of the

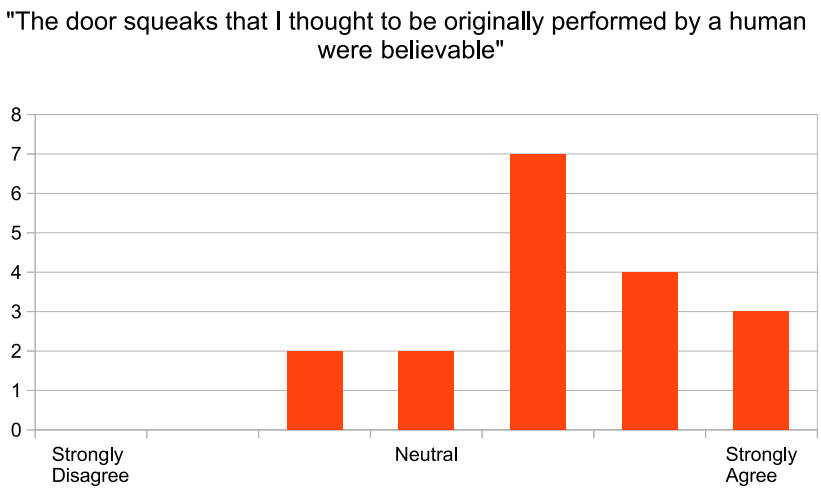


Figure 4.18: Participant ratings of believability for sequences believed to be performed by a human.

| | % Responses correct | Total responses | p-value |
|---|---|---|---|
| Based on recordings | 44.4% | 45 | 0.276 |
| One-to-one | 56.3% | 32 | 0.189 |
| Convergent | 69.0% | 29 | 0.012 |
| PhICL | 55.2% | 29 | 0.229 |

Table 4.8: Percentage of correct listening responses for each category. P-values correspond to binomial tests.

experiment I wouldn't have been able to tell'. Another participant mentioned that while they could 'feel some kind of pattern in the performed sounds' they wouldn't have been able to tell if they hadn't participated in the first part of the study. Some participants also identified specific traits in the sounds that caused them to believe they were performed. One claim was that performed sounds 'seemed over-elaborate', while others commented on the fact that some sequences identified as performed were too long in duration. The latter suggests an unfortunate limitation of the study, in that while performed sequences corresponded to similar classes of doors occurring in the performance study they were not performed within the same narrative context.

Figures 4.18 and 4.19 show that participants generally found sequences to be believable regardless of whether they perceived them as being performed or based on a real recording.

### 4.6.6 Discussion

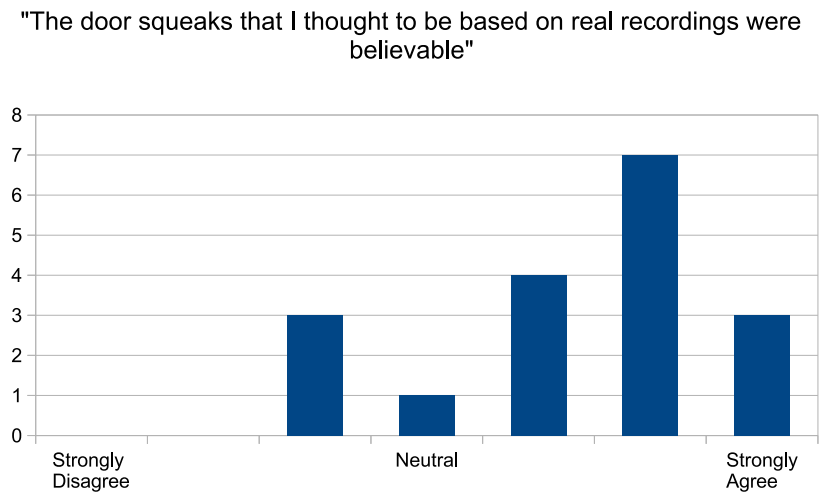


Figure 4.19: Participant ratings of believability for sequences believed to be based on recordings.

The many-to-many mapping received the plurality of positive responses in the questionnaires and also received the highest ratings for repeatability (both other metrics showing no significant differences in scores across interfaces).However, these sounds were also most readily distinguishable as being performed, pointing to potential constraints in levels of behavioural breadth obtainable by the control layer. This suggests that, potentially unlike musical instruments, the reaction of the performer might not be the best (or certainly not the only) metric in evaluating the success of the interface.

The consistent low ratings of the PhICL and comparative success of the many-to-many mapping may seem somewhat surprising considering the intuitiveness and controllability that is often associated with physical control mechanisms. On the other hand, ratings are likely to change after longer exposure to the interface. For instance, the PhICL might yield much higher repeatability ratings after participants have spent more time practising performing sounds using this interface. The only way to map the learning curve of a performance interface is through a longitudinal study (Jordà, 2005; O'Modhrain, 2011). However, it is important to remember that, unlike musical instruments, performable interfaces for computational audio models would require a very steep and low learning curve if they are to be used as part of the design process. The PhICL was based on a natural but difficult interaction (e.g. learning how to sustain a note on a violin can take a long time) and this perhaps contributed to its perception of being the most challenging control layer. It would be interesting to explore more uses of metaphorical behavioural models in the mapping of physical

sensors to timbre-led models, to see whether other metaphorical behavioural models can result in less challenging yet effective interactions.

It is impossible to make a value judgement over which control strategy is better, especially within the constrained scope of this study. The findings above suggest that there can potentially be a lot of freedom in designing control layers. The preference of the arbitrary mapping layer is particular pertinent here.

The most important finding is that, within the scope of this model, performative approaches are indeed a viable option in the rapid creation of new sound effects for a given narrative context. In other words, it is possible to perform behaviours with a physical controller instead of manually programming parameter trajectories – without having an impact on their perceived believability.

It is worth reiterating here that this study is not an exercise in physical modelling or in the matching of existing recordings. Instead, it is an attempt at leveraging human expressivity in a practice that has otherwise required high levels of specialist expertise falling outside the professional scope of a typical sound designer. While a programmatic approach (e.g. in the use of a breakpoint editor) is effective in matching a specific behaviour observed in an existing sound, it is not suitable for exploring and rapidly performing sound effects in response to a specific narrative brief.

A viable direction for future work would be to repeat similar studies with a wider variety of sound models, including non-continuous sounds such as door slams or explosions. The ability to rapidly compose sounds on the fly through performance makes computational audio a much more viable option in conventional media that rely on fixed recordings. This is particularly relevant in theatre, where sounds should ideally be matched to sometimes extemporaneous movements on stage; or in television production, where only limited time and budget is allocated to audio post-production. In the development of the base sound models themselves, design processes that are led by aesthetics rather than physical simulation (such as the four-stage iterative procedure described above) allow designing models for a specific production context; eliminating issues of genericity ('one-size-fits-all' models that become recognisable) and sterile sound quality often attributed with physical models.

Looking ahead, computational audio is likely to become an increasingly crucial element in mediated reality experiences and other media that rely on nuanced interaction. Performative techniques can help design sound effects that are both expressive and malleable while avoiding many of the artefacts and labour involved in integrating fixed recordings into an interactive environment. This, of course, comes at a cost of introducing inconsistencies between the physical action represented through any existing non-auditory channels. While source models provide a simple means to interface parameters from a virtual environment with the sound model, further mediation is required to integrate performed behaviours of a timbre-led model. This will be the

focus of Chapters 5 and 6.

### 4.6.7 Summary

Using a set of four metrics inspired by DMI design - *micro-diversity*, *mid-diversity*, *repeatability* and *believability* - three mapping strategies developed to control a physically-informed model of a squeaking door were evaluated. While an arbitrary convergent mapping strategy had the highest ratings in both quantitative analysis and subjective responses, believability ratings gathered from a listening study suggest that player-centric evaluation may not be enough to evaluate the suitability of a performance interface. While a longitudinal approach would eliminate potential biases associated with the learning curve of the interface (e.g. initial unfamiliarity), the immediacy of a performable sound model is important if it is to be feasibly implemented as part of a larger design process.

The next two chapters shift the focus away from sound model architectures and control strategies, and towards the integration of CGA with moving images. While the model and control strategies presented here offer a promising approach in the performative composition of individual sound effects, means of integrating these performances into an interactive scenario have not yet been addressed. Chapter 7 returns to findings from this study, discussing them in the light of integration strategies and broader implications on CGA and sound design for interactive environments.

# Chapter 5

# Objectives and Apparatus for Studying Performed Sound Synchronisation

The previous two chapters have focused on the design of sound models with the aim of providing a perceptual (as opposed to behavioural) parameter space to the sound designer. In turn, this space was mapped to control interfaces using various mapping techniques. For the case study of the creaking door model this has shown to be an effective approach for performatively generating individual sound effects.

Less attention has been directed towards their eventual integration into a virtual and/or interactive environment, where continuous movement is an increasingly dominant feature. The following two chapters are focused more directly on human factors in sound design for continuous movement (as opposed to single events) and how these can further inform the use and integration of computational audio models into digital media.

Current strategies for designing interactive soundtracks are based on what will be referred to here as an *event-sample* paradigm. Using dedicated software (often referred to as *audio middleware*), several fixed waveforms can be blended, manipulated and triggered at runtime based on events occurring in the virtual scene. For example, a recording of a door creak might be played back once the middleware receives a corresponding event from a game engine. In a more elaborate implementation, multiple waveforms might be employed to achieve greater variety, for example, blending them or varying their playback speed as a function of rotational velocity or door angle. Problems with this approach have been discussed in Section 2.2.3 and include difficulties in providing continuous aural feedback, problems of repetition and a general

lack of nuance in audiovisual synchronisation.

While continuous feedback and non-repetitive audio are clearly important problems to tackle, it is worth considering whether having tighter, more nuanced synchronisation between the sound and image is crucial, or even desired in creating an expressive soundtrack. As noted by Chion (1994) and many other film scholars (see Section 2.3.1), disparity between the sound and image is often *necessary* in order to add further meaning to the audio-visual image. For example, a sequence of a car crashing into a wall might signify an emotionally climactic moment in a narrative, which the soundtrack would seek to underline. In one potential approach to the soundtrack, the initial blast might be unusually loud and dominant in low frequencies, followed by a long tail of debris hitting the floor and fracturing. The exact timings of particles hitting the floor and other secondary visual components may be less important here than maximising the overall emotional or visceral impact of the sequence. Another strategy might be to use silence as an expressive device, rendering only certain elements of the screen and thereby disrupting the flow of the audiovisual image (see Sonnenschein (2001) and Chion (1994) for some interesting discussions on the use of silence in cinema). In another scenario, the scene might have been shot in slow motion, which would require a different approach to the soundtrack altogether, foregoing realism and heightening microscopic elements and movements instead (Epstein, 1985).

The process of audio post-production relies heavily on Foley artists to physically perform elements of the soundtrack. A wide variety of sound producing objects (commonly referred to as *props*) are used to render individual components of the soundtrack. These don't always correspond directly to the visual element of the screen – for example, leather gloves are often used to perform the sound of birds taking flight. Aside from any particular acoustic qualities these props provide the opportunity for artists to dynamically vary the sound, effectively re-enacting the corresponding visual behaviour through performance.

This approach is at odds with current techniques for integrating CGA into a visual or other non-auditory context (see Section 2.2), where individual elements in a scene are often sonified in a bottom-up approach and on a per-component basis. Sound models are normally parameterised to respond directly to physical parameters (for example, *velocity, surface roughness* and *downward force* for a rolling barrel). Other methods employ complex physical simulations solved at audio-rate to synthesise both auditory and visual images simultaneously. In contrast, translating conventional aesthetic approaches would involve designing the model around the expressive function of the sound in its given context, rather than following a generic description of the physical event at hand. The aim of the next two chapters is to explore the implications of such an approach to CGA, through the lens of Foley art.

Section 5.1 provides some additional background pertaining to current integration strategies for generating real-time soundtracks for interactive environments. These include *one-to-one* mappings of physical data to sound models, the event-sample paradigm that is currently widely adopted in the games industry and potential approaches for computational models based on parameter sequences proposed in the previous chapters.

Section 5.2 presents results from a questionnaire conducted with professional Foley artists exploring aspects of performance and synchronisation deemed important in a conventional motion picture context.

Finally, Section 5.3 describes an experimental environment that was developed to study sound-image relationships as performed by professional Foley practitioners using physical interfaces driving computational models.

## 5.1 Integration Strategies

The previous chapters have explored how human performance can be leveraged to generate expressive audio on the basis of discrete sound effects. However, they have not considered the role of sound over larger scales of time, or in response to a continuous flow of moving images (as opposed to single events).

A brief overview of current techniques for synchronising sounds to interactive images - a process commonly referred to as *integration* (Collins, 2008) - will be presented here.

### 5.1.1 Event-Sample Paradigm

Currently, the prevalent approach for developing interactive soundtracks in the games industry is the use of so-called audio middleware to manage the playback and processing of large amounts of sound files based on states and events occurring within a game. For example, if a character in the game jumps, this either triggers or is triggered by a 'jump' event, which the software can be programmed to respond to by playing back a corresponding sound file (see Figure 5.1). A more elaborate implementation might trigger a particular combination of sound files depending on the intensity of the jump, the character's state, or any other available variables that are meaningful to the sound designer.

Mapping sound effects to selected events provides players with consistent and reliable feedback, helping the familiarisation process with the virtual environment (Vachon, 2009). However, this approach can often lead to a degradation in quality and immersion when repeated events result in repeated sounds (Mullan, 2010). Furthermore, with the increased capabilities of runtime physics engines and the introduction of continuous control over the environment (e.g. in mainstream virtual

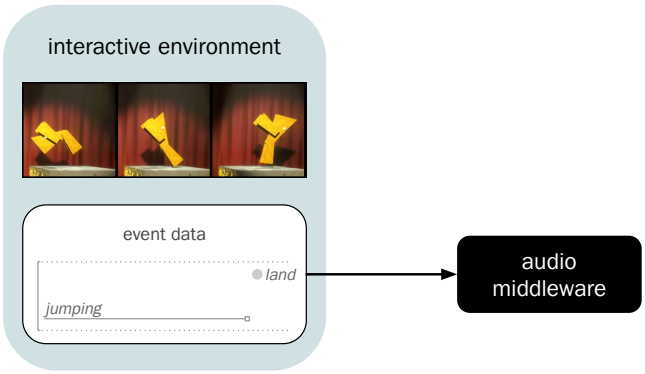


Figure 5.1: Event-based sound integration common in the games industry

reality applications), event-based audio integration techniques can easily fall short of providing reliable feedback that follows continuous movement while simultaneously maintaining a high level of sound quality. Two strategies can be applied to mitigate this problem (both of which are common techniques in the use of audio middleware (Collins, 2008)).

The first is to randomise the playback and mixture of different assets in order to make the repetition of pre-processed sound effects less obvious and therefore provide feedback that is perceived as being 'more organic' (Vachon, 2009).

The second strategy is to categorise events on a finer grain of detail, taking into account their context within the wider narrative (e.g. how often this event has occurred before) or other states in the virtual environment (e.g. the condition of a vehicle changes the sound of the engine starting up). While this is currently seen to provide the most efficient solution to composing an engaging interactive soundtrack there are several shortcomings to this approach:

- The number of required audio assets grows with the introduction of new game content
- Randomisation, while less fatiguing, is still monotonous
- Congruence to movement (or *kinesonic congruence* (Collins, 2013)) is lacking due to obvious constraints of static assets

### 5.1.2 Physics-Based Integration

One of the greatest perceived benefits of physics-based sound models is that they bypass many of the constraints of the event-based approach applied in audio middleware by interfacing directly with physical actions inside the virtual environment (e.g. as

produced by a physics-engine or through first-person interaction). Thus, the sound of an object scraping across a surface can be produced by feeding a sound model with a stream of velocity and force values extracted from the sonifiable system. Alternatively, the images and soundtrack could both be produced by the same physical simulation (Verron and Drettakis, 2012). In both cases, visual and sonic behaviour are strongly coupled, and could therefore be regarded as the virtual equivalent of using direct sound recorded on a film set.

The concept of the discrete 'sound effect' has reduced relevance here, as does the notion of such sound effects being triggered by discrete events within the virtual environment. More crucially, it is very difficult to introduce any stylistic intervention beyond adjusting the fixed parameters of the model. In other words, the way that the model reacts to physical parameters can be altered, however the mapping between physical states and model parameters results in a strict consistency between sound and image across longer durations of time. Just as the previous chapters have likened numerical models of sound to 'black boxes' that leave little opportunity for the sound designer to intervene on a timbral level, the larger scale temporal behaviour of the model in this integration strategy is similarly opaque, albeit very precise in its synchronisation.

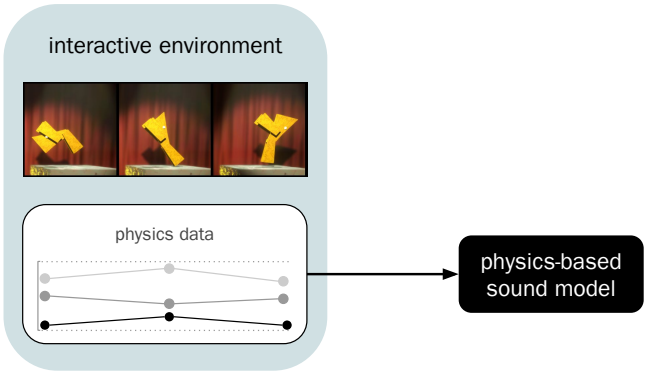


Figure 5.2: One-to-one mapping of physics data to a physics-based sound model.

### 5.1.3 Event-Based Triggering and Blending of Pre-Composed Behaviours

The event-based integration strategy (see discussion of the *event-sample* paradigm above) could be applied to performed behaviours for computational models, where behaviours constitute streams of parameter values for a given model (as introduced in Chapters 3 and 4). Parameter data representing individual performances can be

combined or manipulated *before* being fed into the model, giving rise to many new opportunities that are unachievable with sample-based assets (see Figure 5.3). For example, two sequences performed on the timbre-led model of the creaking door described in Chapter 4 could be blended as a function of a dynamic parameter from the virtual environment.

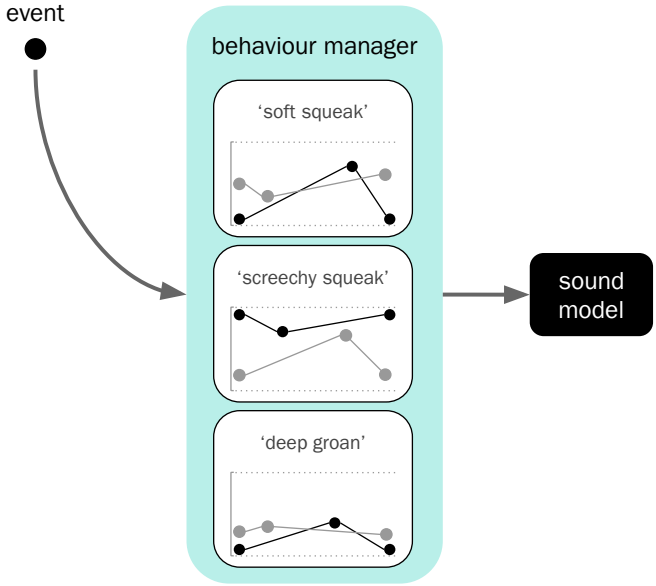


Figure 5.3: Event-behaviour paradigm applied to computational models

As discussed in Chapter 3, maybe the most immediate gain from doing this in comparison to a sample-based approach would be a significant drop in memory requirements per audio asset, especially if the parameter trajectories can be encoded into keyframe (or 'breakpoint') data. The event-based nature of this approach means that sound behaviours can be composed on a bespoke basis, overcoming limitations of physics-based CGA discussed above. Some of these integration techniques were explored in the *FoleyDesigner* project (see Appendix D), which uses *meta-parameters* to playback and interpolate various performances in real-time using a game engine. Benefits and limitations of this approach are discussed in Chapter 7.

### 5.1.4 Between Events and Continuous Movement

By accumulating performances of a computational model and treating them as conventional assets one could argue that the interactive soundtrack can enjoy the best of both worlds. Audio assets occupy much less memory, kinesonic congruence may be more easily achieved due to the continuous nature of the parameter space (e.g. fewer

artifacts arise from blending and modulating playback speeds) and the sound model can be designed according to the priorities of the production (e.g. timbre) instead of being constrained by physical processes.

However, with the increasing use of physical simulations of motion and first-person interaction (in both games and virtual reality applications) the event-based workflow falls short of being able to follow continuous and, oftentimes, unpredictable movement. In other words, while the event-based approach is likely to result in a soundtrack with inadequate levels of responsiveness, the physics-based approach leaves insufficient room for creative intervention in the synchronisation of sound to the image.

In motion picture sound Foley artists are typically employed to performatively synchronise sounds to continuous movements on the screen. This process makes it possible to obtain a soundtrack that follows continuous movement while simultaneously leaving opportunities for exaggeration, emphasis or other forms of stylisation. In the light of performative sound synchronisation, both physics-based and event-based approaches can be challenged on the basis of common aesthetic devices in the composition of audio-visual images for film (Chion, 1994) and gestural expressivity and prosody.

Tuuri (2009) advocates a heightened emphasis on prosodic elements in sound design by incorporating gestural attributes rather than relying on purely symbolic meaning (echoing the perspectives of Leman (2012) on kinaesthetic meaning transferred through embodied interaction). This has generally been approached on an object-by-object basis (where 'object' can be understood in the Schaefferian sense of the 'sonic object'). Borrowing terminology suggested by Godøy (2010), the focus of the previous two chapters has been constrained to the *micro* timescale of gestural-sonorous objects. Considering the *macro timescale* ("[...] that of larger-scale forms such as whole sections, movements, and whole works, consisting of the concatenation of successive sonic objects") reveals how gestural movement can carry meaning across larger sequences of sound objects. Assigning sound effects on an event-by-event basis precludes the sound image from stylistic divergences over larger scale sequences, resulting in temporally disconnected 'snapshots' of sonic movement rather than a meaningful whole.

Though less focused on gesture per se, a direct parallel to this argument can be found in film sound theory, particularly in frameworks developed by Chion (1994). Chion's concept of *audio-visual phrasing*, which likens the relationship between sound and image to a sort of *counterpoint*, relies entirely on the soundtrack's ability to evolve consistently over a prolonged period of time. For example, the concept of *temporal vectorisation* is defined to occur "when a number of sound elements and/or visual elements are superimposed and constituted in a way that allows the spectator to anticipate their crossing, meeting, or collision just ahead".

The remainder of the work presented in this thesis explores potential future integration strategies for computational audio by analysing performed soundtracks to virtually represented movement. Foley art represents a fruitful field to study in order to explore the types of synchronisation strategies employed in generating expressive soundtracks.

The next section presents results from a survey of Foley artists addressing a variety of themes ranging from interaction with sound-producing objects to priorities and techniques in the synchronisation of sound. Observations from this survey informed the development of an experimental apparatus, presented in Section 5.3, for the detailed study of soundtracks performed by a small group of Foley artists.

## 5.2 Survey of Foley Artists

An online survey was conducted with twenty-five Foley artists and sound designers with professional Foley experience to assist the process of designing an experimental environment, and to inform later analysis of the obtained data and findings. Six of the respondents took part in the final practical study discussed in Chapter 6.

The survey consisted of a mixture of Likert-scale and free-response prompts within the following categories:

- General information on previous experience and demography
- Use of sounding objects (*props*) to perform sound to moving images
- The relationship between sound and image in performed Foley tracks
- Experience in and opinions on the use of novel technology for performing Foley
- Experience in performing Foley for interactive media (if applicable)

The survey consisted of an online questionnaire targeted at Foley artists and professionals in audio-production with significant experience in performing sound to moving images.

### 5.2.1 Respondents

Twenty-five participants aged between 20 and 65 took part in the survey.

Thirteen respondents had more than ten years of experience in Foley art or sound design involving substantial amounts of Foley work. Eight participants had between two and five years of experience and four had between five and ten years of experience.

Foley experience ranged across various types of media including short and feature-length film, television, commercials, animation, games and theatre.

### 5.2.2 Structure of Survey

While a variety of broader themes were addressed in the survey, this overview will focus on two main sections that are directly relevant to issues of asynchrony and performance introduced above. The first section was based on the interaction with physical objects (referred to as *props* (Ament, 2009)) in the synchronisation of sound, focusing on issues of navigation and exploration, re-use of props and perceived distinctions between timbre and behaviour. The second section focused on strategies and techniques in constructing soundtracks to the moving images, including temporal and physical divergences from depicted movement and narrative context. The full questionnaire is presented in Appendix B.1. Supplementary illustrations of results that will be referred to throughout this section can be found in Appendix B.2.

The following two subsections present an overview of key observations made in the above two categories of interaction with physical objects and the relationship of sound to the moving image, each beginning with a brief description of its relevance to the issues addressed in Section 5.1.

### 5.2.3 Performative Interaction with Physical Objects

The relationship between the performer and computational sound models (including any physical controllers and digital control layers involved) has been addressed from a technical point of view in Chapter 4. Here it is contextualised more directly within the existing cultural practice of performing sound in time to a moving image (Foley). Literature on Foley art is somewhat limited, and the deepest insights are often found in testimonies from artists (e.g. Ament (2009)). The following aspects are most commonly addressed in existing literature:

- Inventive use of objects to recreate sounds (e.g. ballet-shoes for birds, sheets of metal for thunder, etc.)
- Audio-visual *synchresis* (Chion), i.e. the common sensory fusion of audio and visual channels exploited extensively in Foley.[1]
- Workflow and the production infrastructure - the role of editors, Foley artists and mixers, time management and project management.

There is less focus on the actual interaction between the Foley artist and the sounding objects (props), and the activity is often instead referred to as a 'dark art'. This survey, on the other hand, has revealed some interesting perspectives on the use of physical objects to produce sound.

[1] This has also been studied in psychophysics, the McGurk effect being a widely known example (McGurk and MacDonald, 1976).

### Re-use of Props

Respondents were asked to estimate the percentage of props that they normally would have encountered for the first time when working on a project, and of those that were used regularly across different projects. The average percentage of props encountered for the first time was 26.84% (s = 18.7%). The average percentage of props regularly re-used across projects was 66.32% (s = 22.25%).

### Exploration

When prompted whether they spend a lot of time exploring ways of handling props and exploring timbral characteristics, respondents responded positively in both categories (with a slightly stronger response to ways of handling props; see Figures B.2 and B.3). The overwhelming majority of respondents were often surprised with the sounds a given prop is capable of producing, despite regularly re-using props across projects (see Figure B.5. All but two respondents stated that they pay a lot of attention to timbral nuances of the sound while performing with a prop (see Figure B.10).

### Challenges in Handling Props

There were mixed responses to prompts regarding perceived difficulties in performing with props. While many respondents responded in agreement to the prompt "I often get frustrated with some props because they are difficult to perform with", there were more mixed responses to the prompt "I often get frustrated with some props because they are unable to produce the sound I want them to" (see Figures B.4 and B.6). The overwhelming majority of respondents stated that they find it easy to recreate the same take using the same physical object (see Figure B.1).

### Embodiment and Timbre

Perhaps unsurprisingly, there was strong disagreement to the prompt "the human performed aspect of Foley causes the dubbed soundtrack to be less than ideal". About half of the respondents disagreed with the prompt "I focus more on on-screen (or on-stage) action than the timbral nuances of the sound" (see Figure B.7). This is consistent with the observation that most respondents pay a lot of attention to timbral nuances. Most respondents agreed that "handling props is much like playing a musical instrument"(see Figure B.8).

### Further Observations and Discussion

Many participants had difficulty differentiating between 'good' and 'bad' props, emphasising that context plays a fundamental role in the perceived quality and suitability

of a prop. One respondent commented "it's the way you work with a prop" rather than the prop itself, which usually depends on context and "what works for a given shot". This perspective was shared across other respondants, as illustrated in the following statements:

- *In my belief, performance, sound quality, and sync are all interconnected. The right prop performed wrong is simply not the right prop. You do whatever it takes, change the performance, change the prop to get the "right sound" whatever that may be.*
- *It's hard to define a prop as good or bad because that means you first have to put it in a certain context. Is that large metal bucket bad for a bucket sound, maybe. But, it turns out it is great for rolling on its side and sounding like a grenade... Most of the time a prop underperforms where I think it will excel and excels in an area I never thought of. It is learning what an item can and can't do that is valuable.*

These statements paint a picture of Foley props that stands in stark contrast to musical instruments. First of all, the concept of virtuosity and learning curve doesn't seem to apply so much to individual props as it would to performing a wide variety of objects well within a diverse range of narrative or stylistic context. While most respondents stated that they reused existing props across projects more often than they would encounter new objects, they were often surprised by the range of sounds that could be produced.This can be assimilated to Jorda's concept of 'macro-diversity' (Jordà, 2005), which is used to describe or measure the variety of stylistic contexts a given musical instrument can be played within. In Section 4.4.4 this was discussed in light of the sound-source relationship, proposing the alternative term *source appropriability*.

On the level of timbral manipulation it also seems that *range* and *repeatability* (as discussed in the previous chapter) are deemed to be important factors. Participants complained about props that had a narrow dynamic range and that produced unpredictable output upon handling. Ergonomic factors also seemed to play a role:

- *A bad prop is clumsy and doesn't allow much in the scale of gentle to aggressive movements.*
- *Bad Props: One hit wonders (although there's a place for them), not expressive, random, hard to use, boring resonance, hard to place on surfaces.*

Despite Foley artists relying on a very large arsenal of props that are often used in entirely new ways and contexts, there is still a strong sense of aesthetic sensitivity

and experience that could be compared to the concept of virtuosity in musical performance. This appears to take on two dimensions: physical believability (including timing and material congruence) and the projection of mood and/or emotion:

- *Sometimes a prop might not work for a particular instance, but on another project will be PERFECT! I like to say "There are no bad props, only bad Foley Artists!"*
- *We need to perform a sound to match the picture. We also perform the prop to capture the mood or energy of the action (hence the term "perform") [...] Every sound we create is an aesthetic choice!!*
- *The baseline requirement of a good Foley artist is the performance. Sound quality on the mix end and sync is one thing, but understanding the emotion and the performance is the literal part of the job. Once someone figures out footstep technique, for example, it matters more to nail the nuances of the performances.*

It would seem that many parallels can be drawn to free musical improvisation, where virtuosity often does not come from mastering a single instrument or playing style, but rather from being able to project meaning within diverse contexts and very often using a multitude of different instruments or objects and experimental playing styles.

### 5.2.4 Performer and Moving Image

Viewed broadly, the relationship between the sound artist and the moving image would at very least encompass the entire discipline of sound design. The sound-image relationship can be studied on various temporal or *horizontal* scales: the full duration of a film, the composition of shots forming a scene, a single shot, individual actions within the shot and the micro-gestures (or *chunks*) that these actions can be separated into. There is also as much scope in the *vertical* composition of a soundtrack, ranging from ambience and music to dialogue and diegetic sound, and the ways in which all these elements are combined into the final mix.

While these are all important factors in the composition of an audiovisual scene, the scope of this thesis is limited to discrete sound models and strategies for coupling them to visual movement. Therefore, the object of interest here comprises only a small fraction of this huge space. On the vertical dimension the primary focus is on diegetic, concrete sounds that are associated with visible on-screen objects. Horizontally, the focus is on the actions of these objects, their intrinsic shapes and structures, and how these might be affected by the narrative progression of the image. Foley art therefore represents an ideal professional practice to study as artists generally work on an object-by-object basis when synchronising sound and usually without hearing

any of the rest of the soundtrack during the performance. Although the Foley track to any given object is always performed within a much larger context of a narrative and stylistic brief, the fact that sounds are performed in isolation makes Foley an adequate lens through which to investigate the aesthetic dimensions of synchronising movement with the virtual counterpart of a 'prop' (i.e. a sound model). As in the previous section, a concise set of observations from the questionnaire will be presented, followed by a broader discussion taking participant responses into account.

#### Temporal Consistency

Most participants agreed with the prompt "I often sacrifice temporal accuracy in order to improve other aspects of the sound" (see Figure B.17). This is maybe consistent with a further observation that about half of the participants rely on post-editing to achieve good timing (see Figure B.15). Most respondents didn't use intentional temporal anticipation or delay in relation to the moving image for aesthetic purposes (see Figure B.16).

#### Physical Consistency

The majority of respondents stated that their performed sounds are usually exaggerated versions of the corresponding on-screen actions (see Figure B.12). About half of the respondents try to maintain consistency across similar actions, even when making physical exaggerations (see Figure B.13).

#### Narrative Context and Briefing

The majority of respondents agreed that narrative and aesthetic details play a big role in the way Foley is performed (see Figure B.19). It is worth also reiterating here the above-mentioned observation that the vast majority of participants strongly valued the human-performed factors in the dubbed soundtrack.

#### Number of Takes and Grounds for Rejection

Respondents were asked to estimate how many takes, on average, are normally performed for a shot or scene before being satisfied with the result. Most respondents required between one and five takes while four respondents claimed to require more than five (but never above ten).

Respondents were also asked to rank their reasons for rejecting a performed take, selecting between 'timing errors', 'timbral detail', 'ergonomics', 'technical faults' and 'other'. Timbral detail was the top-ranked reason for rejecting takes, followed by timing errors (see Figure 5.4).

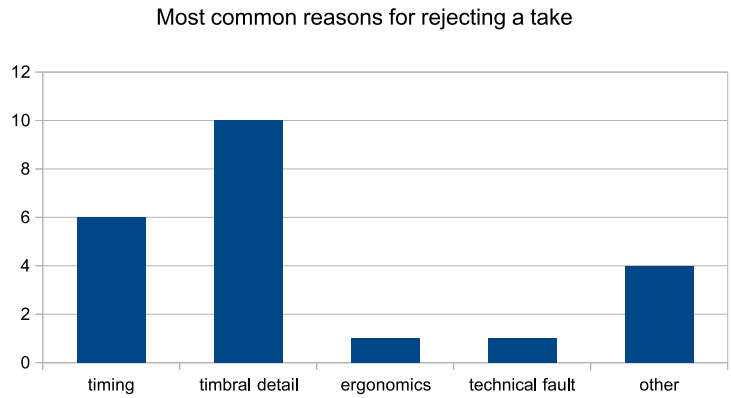


Figure 5.4: Top rankings of most common reasons for rejecting takes.

### Further Observations and Discussion

Several respondents were keen to point out two contrasting objectives in the process of performing Foley. On one hand, the performance needs to portray and heighten the emotions and actions represented by a character within the narrative and stylistic context that it exists in. On the other hand, the Foley artist has a duty to maintain a level of believability and authenticity, which pertains to material properties of the prop and, maybe most pertinently, to the quality of synchronisation to the on-screen movement.

- *Foley serves four masters at all times: creating a believable, grounded world; expressing the determinations of the character; expressing the emotion or genre expectations of the moment; and BUDGET.*
- *Foley during a transitional moment in the film MUST be both expressive of the emotion/meaning needed in the script and believable to the audience as coming from the emotion created in that scene. However, today's audiences need "authenticity" also (unless the genre's audience expects surrealism at that specific plot point).*

This dual focus on sonic congruence to the image on one hand and expressivity on the other is maybe what sets Foley apart from most other fields of performed sound. Indeed, the way in which aspects of synchronisation and expressivity (or 'added value' in Chion's words) interact in the soundtrack marks a fundamental conflict that lies at the heart of the research problem: in order for a soundtrack to be expressive it needs to diverge from the physical reality of what is represented on screen. As discussed above, the soundtrack can diverge from the physical reality on both *vertical* (e.g. 'intensity'

or other physical qualities of a sound and its material properties) and *horizontal* (e.g. timing and punctuation) dimensions. While these two axes are a technically convenient way of approaching the problem (and eventually comparing performed soundtracks to physical reference data), many participants in the survey were hesitant to describe their craft in these terms, typically using much more technically ambiguous vocabulary:

- *The footsteps had the right weight, the cloth movements could be performed in an expressive manor that emphasized the action and sound design. Furthermore, Foley can be expressive through its texture, timbre, and arrangement.*
- *There is beauty in a Foley track when it is so clean that it gives the illusion of it not being present yet the characters magically gain dimension.*
- *The baseline requirement of a good Foley artist is the performance. Sound quality on the mix end and sync is one thing, but understanding the emotion and the performance is the literal part of the job. Once someone figures out footstep technique, for example, it matters more to nail the nuances of the performances.*

Some participants, on the other hand, went into some more detail about when and how they might let the soundtrack diverge from the film's timing:

- *If moving a Foley gesture forwards or backwards in time helps it work better with the other sonic elements, I will sacrifice temporal consistency. And generally, if it isn't something that needs to be tightly in sync with picture, the audience will never notice.*
- *Lack of sync creates unease in the viewer when performed correctly.*
- *More often than not the choice comes down to if you want [the performance to be] slidy or more patty percussive or scraping or foldy [...] whatever movement best 'sells' the action. We have to make a lot of things sound like they're in motion.*
- *Animation and video games. Since those are usually almost perfectly edited frame for frame, or in the case of gaming, only has let's say 3 animated instances of a body fall, we're there to add the human element to trick the eye and ear into seeing more organic performances.*
- *To bridge edits from the A side to the B side, to match animal footsteps (horses) or to add a comic or animated quality.*

In summary, Foley artists have a clear intention of re-interpreting, magnifying and curating elements of on-screen movement when performing live synchronisation.

The precise ways in which the sound deviates from the image is difficult to deduce from comments alone and, understandably, is not necessarily something that the artists themselves are aware of. From a technical point of view, the soundtrack can deviate in terms of *horizontal* synchronisation and *vertical* representations of forces and material. In addition to this, each Foley artist will approach a scene differently, based on their own previous experience and interpretation of the movement and visual context. On the other hand, conventions in post-production, cultural conditioning and other factors are likely to result in commonalities across practitioners in the ways they cause the soundtrack to deviate from the image. One of the aims of the final study is to try and delineate a set of stylistic devices based on recurring interpretations of physical movement.

### 5.2.5 Summary

The relationship between the soundtrack and the image (within the context of a single object's continuous movement) lies at the core of what the following study seeks to shed light on. While testimonies from Foley artists provide some insight, a more precise means of analysis is required in order to establish how intentional deviations between sound and image manifest themselves within a narrative context and, ultimately, in the sonification of physical movement using computational sound models. The experimental environment outlined in the remainder of this chapter was developed with all of the above considerations in mind.

Considering the wealth of possible imagery, types of sounds, narrative contexts and stylistic interpretations, a single study can only scratch the surface of the entire range of aesthetic decisions that are made in the human synchronisation of sound. Therefore finding a definitive and precise integration strategy that reflects these decisions falls well beyond the scope of this research project (and may even be an impossible task). Nonetheless, focused analysis of a carefully chosen case study based on both qualitative and quantitative observations is the logical next step that can open the door to increasingly informed and calculated investigations. The remainder of this chapter follows a technical overview of the experimental environment that was designed to facilitate this study on a component-by-component basis.

## 5.3 Technical Overview of the Experimental Environment

### 5.3.1 Objectives

The central aim behind this final phase of research was to identify stylistic strategies that can be employed in the generation of computational soundtracks to visual phys-

ical movement. As discussed above, asynchronicity between the sound and image plays an important role in imparting intentional meaning through the soundtrack. From an interactional angle, frameworks of *co-articulation* (Godøy et al., 2009) and *prosody* (Tuuri, 2009) suggest that expressive qualities of sound emerge through physical movement and inherent hierarchical temporal structures.

This study therefore leverages expertise from the field of Foley artistry in order to evaluate how performative strategies can aid the intentional stylisation of computationally generated soundtracks. On one hand, conventions and techniques applied by Foley artists may transfer directly to the integration of sound models, on the other hand the technological constraints imposed by the study may highlight shortcomings and opportunities in CGA and the interactional strategies employed.

As the integration dilemma identified in Section 5.1 applies most pertinently to motion that is complex or emergent and emotionally evocative, the graphical material chosen and developed for this study was a physics-based animation of a simple anthropomorphic figure. Aside from the naturally emergent qualities of animation driven by physical principles, it has the added benefit of affording a direct mapping to physically corresponding sound models, thus offering a means to produce a plausible reference soundtrack. On the other hand, a parametric state-machine system interacting with the physical simulation provides the character with natural anthropomorphic movement. This also provides the scene with a discrete event-based reference that would be more commonly used in current sample-based integration strategies.

A physical interface was developed that facilitated the performance of impact, scraping and squeaking models. Unlike the previous study these models retained their behavioural abstraction layers in order to obtain a plausible reference track. Some corresponding behavioural parameters were exposed to help participants fine-tune the interaction – see Sections 5.3.5 and 5.3.6. Performance data could then be compared directly with the physically generated data to highlight specific intentional divergences.

By giving professional Foley artists, dubbing editors and sound designers the ability to perform the soundtrack to game-like animation sequences using a set of computational audio models, the resulting dataset could later be used to study the structure of each subjective soundtrack in relation to:

- the physics-based mapping of sound models to movement on a one-to-one basis (by comparison to a corresponding reference track)
- an event-based approach (informed by events and states extracted from the animation)
- the subjective structure and categorisation of the soundtrack (through in-situ testimonies by the participant)

### 5.3.2 Summary and Criteria of the Experimental Environment

In order to facilitate this study the following components were deemed necessary in the experimental environment:

- an evocative physics-based animation from which physics data can be extracted
- computational sound model(s) capable of sonifying this data with an adequate degree of believability
- a physical interface for performing the same computational model(s)
- a means of easily performing and recording model parameters in synchronisation to the moving image with an adequate degree of precision and resolution

These were developed in the above order and will now be introduced briefly. A more detailed technical summary of each component is provided in the following sections.

#### Animation

A procedural physics-based animation system was developed in parallel to this study. The animation is of a wooden anthropomorphic figure consisting of five separate wooden pieces (or limbs). A state machine allowed transitions between various states such as jumping, running and walking, which in turn were achieved by means of physical simulations of mass-spring-damper mechanisms on each limb.

While the visual material was a static (or 'linear') film, it was essential that physical data could be extracted, pertaining to the depicted movement. While following a prescribed narrative, the animated character was designed to move according to a set of discrete states or events, allowing soundtracks to be approached in light of a general structure of movement that can be found in games and other interactive scenarios (e.g. walking, jumping, running, and so forth).

#### Sound Models

The sound models were designed with the basic requirement of sonifying the most immediate physical movement of the animated object. The choice of sound models followed the most likely choice that would be made based on a one-to-one mapping of physical movement to sound. This allowed comparisons to be made to an objective reference track, while constraining the audible material to a sound palette that was manageable within the scope of this study. In the case of the animated character described above the most direct sound mapping was based on the impacts between each wooden limb and the floor. In addition to hitting the floor limbs were also able to scrape along the surface. Existing impact and scraping models were implemented

on a per-limb basis, with individual size and resonance parameters. Finally, a sound model for producing squeaks (based on a simplified version of the creaking door model outlined in the previous chapter) was implemented to sonify the average angular velocity of all the limbs.

#### Physical Interface

The main criteria of the physical interface was for the output to closely resemble the physical data extracted from the animation system: impact force, scrape velocity and angular velocity for squeaks. In order to stay as close as possible to the physical parameters five force and touch-sensitive sensors were built (allowing participants to sonify each individual limb if they so desired) in order to perform impact and scraping sounds. An additional interface resembling a wooden crank or corkscrew was built to control the squeaking sounds. The design of the interfaces was based on *enactive* design principles Essl and O'Modhrain (2006) whereby physical movements required to produce sound were related to the sounds outputted by the system. The aim of this was to ease the familiarisation process with the interface while maintaining similarity between the performer's gestures and the physical movements observed in the animated scene.

#### Synchronisation Workflow

The final versions of the interface and sound models were implemented using Bela (Moro et al., 2016), which provided very low action-to-sound latency while capturing sensor data at a high resolution and fast sampling rate. Sensor data was encoded as audio streams fed into a digital audio workstation (REAPER[2]), which was used to record, playback and edit performances in a way that was familiar to the targeted participant group of professional sound designers, dubbing mixers and Foley artists.

### 5.3.3 Physics-Based Animation of an Anthropomorphic Figure

A parametric physics-based animation system was developed in parallel to this research project. The animation system renders a wooden anthropomorphic figure constructed out of five separate meshes representing two legs, two arms and a head (see Figure 5.5) and was implemented in the Unity game engine[3]. Each limb is driven using simulated physical forces (based on Unity's internal physics engine) and relies on a state machine that interpolates between different types of actions. Actions comprise of locomotion, jumping, rotating, leaning, looking and 'ragdoll' behaviours. The

[2]http://www.reaper.fm
[3]http://www.unity.com

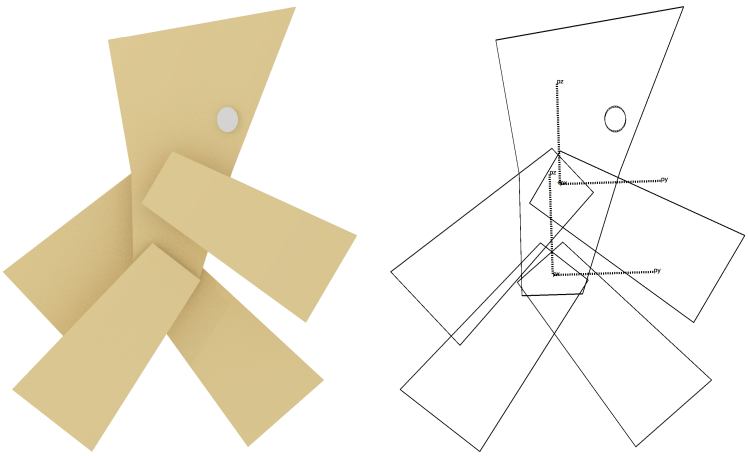

Figure 5.5: 3D model used as the basis for the physics-based animation. Outlines of individual meshes and rotational joints (constrained to forward-facing axes) are shown on the right.

system can be controlled using a game controller, allowing animation sequences to be authored according to a screenplay by manually piloting (or puppeteering) the figure. While the figure is represented in three-dimensional space, the majority of its movements (aside from rotating) are constrained to two dimensions.

Unlike in conventional animations (which typically use a mixture of physics, inverse kinematics and manually scripted movements), extracting accurate physical parameters such as velocities and forces is a trivial task when using an entirely physics-driven animation. While the overall framework for blending animations can be fairly complex, the animator is able to author custom behaviours by programming and tweaking sets of 'poses', setting the corresponding target angles and weights of each individual limb. As the character consists of just five limbs this process is fairly straightforward and the resulting movement strikes a balance between evocative anthropomorphic behaviour and a low bandwidth of physical output data.

### Physics-Based Animation of the Character

The wooden figure's five limbs are connected together using 'spring joints' (an internal object in Unity's physics engine). A spring joint joins together two rigid bodies, one acting as a 'parent' to the 'child' object. The child rigid body is then able to rotate along a prescribed axis in relation to its parent. The spring joint has a variable target angle (of the child body relative to its parent) which it will try to reach based on an internal mass-spring-damper simulation that is configured using corresponding *spring* and *damper* coefficients. As illustrated in Figure 5.5, the 'arm' and 'leg' bodies are

| Name | Type | Range | Description |
|---|---|---|---|
| Weight | float | [0 - 1] | Weighting factor applied when blending states. |
| Target Angle | float | [0 - 360] | Target angle for spring joint. |
| Spring | float | [0 - inf] | Spring coefficient for spring joint. |
| Damper | float | [0 - inf] | Damper coefficient for spring joint. |
| Additional Force | Vector2 | <[-inf - inf], [-inf - inf]> | Additional directional force to apply to limb. |

Table 5.1: Data structure used to control each limb's movement.

connected to the head, meaning that all angles assigned to the corresponding limbs are relative to the head's orientation. Another spring joint is attached between the head and an invisible (or empty) parent object, which is required in order to control the figure's overall orientation and additional head movements. In addition to the spring joints, all limb movements are affected by a global downward gravity force.

Each limb is controlled by a data structure (see Table 5.1) which is updated on every frame of the physical simulation (60Hz). An *action* is defined as a class containing a *process* subroutine which calculates an instance of this data structure based on a set of instructions. For example, a walking behaviour can be generated by cycling through a pre-defined set of poses, testing whether limbs have reached their corresponding target angles before moving on to the next pose. This architecture simplifies the process of adding custom actions by means of inheritance.

### Action Types and Blending

In order to achieve natural movement the animation needs to be able to blend between multiple actions, such that the figure can be 'walking' and 'looking' at the same time, or steadily slowing down to a halt when it shifts from a 'running' state to an 'idle' state. This is implemented by processing multiple actions simultaneously and chaining their outputs together, which are in turn scaled by a *weighting* coefficient. The actions outlined in Table 5.2 were designed in order to create the animation (based on the screenplay presented in Appendix C).

### Piloting the Animation

Input from a game controller was mapped to control global variables that in turn control weightings and other parameters specific to particular action types (such as 'lean direction', 'preparing to jump' and 'walking speed'). The game controller could

| | |
|---|---|
| Idle | Default idle state in which figure stands upright |
| Walk (normal) | A slow and confident walk ('strut') in the current walking direction |
| Walk (sad) | A sad walk ('shuffling') in the current walking direction, slightly limping with head turned towards floor |
| Run | Running ('striding') in the current walking direction |
| Walk (backwards) | Similar to *Walk (normal)* but shorter steps and opposite to the current walking direction |
| Lean | A state in which both arm angles are set according to global 2-D *LeanDirection* vector |
| Look | A state in which the head orientation is set according to a global 2-D *LookDirection* vector |
| Jump | A jumping action consisting of multiple internal states (preperation and in-air). While in-air, body orientation is controlled using *LeanDirection* vector. |
| Rotate | Rotate entire figure to face new walking direction. |
| Limp / Ragdoll | All spring joint coefficients are set to 0 causing entire figure to collapse under force of gravity. |

Table 5.2: Action types designed for animation based on screenplay.

| | |
|---|---|
| Left Analog Stick (horizontal axis) | Walk/run speed and direction |
| Left Analog Stick (vertical axis) | Look direction |
| Right Analog Stick (2-D) | Lean direction |
| 'X' button | Initialise jumping action |
| 'Square' button | Trigger limp state while pressed |
| 'L1' button | Trigger walking state |

Table 5.3: Mappings for a Sony *DualShock 3* controller used to pilot the figure.

then be used to 'puppeteer' the figure according to directions devised in the screenplay. Controller mappings are shown in Table 5.3.

#### Extraction of Physics Data to Drive Sound Models

Physics data extracted from the animation is used to drive three types of sound models: *impacts*, *scraping* and *squeaking*. Each limb is assigned its own impact and scraping models, while the average absolute angular velocity of all limbs drives the squeaking model. The sound models (described below in Section 5.3.5) are physically-informed (Cook, 1997) and react to a stream of physical parameters calculated at a fixed frame-rate of 60Hz. In order to obtain accurate output from these models physical parameters of *force* and *velocity* need to be extracted from the simulation. *Force* corresponds to the total absolute collision force between the rigid body corresponding to a limb and the surface it has collided against (note that the figure is constructed in

a way that limbs cannot collide against each other). *Velocity* corresponds to the magnitude of the *tangential* velocity vector between a limb's rigid body and the surface that it is in contact with.

As the whole animation system is based on Unity's internal physics engine, all of the information required to deduce these parameters is available. For each collision between rigid bodies detected by the physics engine the following data can be obtained which can in turn be used to derive the required parameters: a vector describing the total relative velocity between the rigid body and the object it has collided with, a set of vectors describing the collision normals for each contact point corresponding to the collision, the velocity of the rigid body and the angular velocity of the rigid body. Unity's physics engine also provides callback functions for frames at which a collision has just occurred, frames during which the collision contact (or set of contacts) is still in contact with the colliding object and frames at which two colliding points are no longer in contact.

For each contact point at the occurrence of a collision, the dot product is taken of the relative velocity between the two colliding objects and the contact normal. These values are then summed and scaled by the mass of the rigid body in order to obtain the total collision force. The tangential velocity of the rigid body along the surface can be obtained by taking the magnitude of the relative velocity between the two colliding objects and accounting for any false displacement caused by angular motion.

Average angular velocity to drive the squeaking model is calculated by simply summing the magnitudes of angular velocities for each limb. While this is not the most accurate physical representation this parameter was chosen in order to constrain the soundtrack to only a single friction model, rather than having one for each individual limb.

Each limb has been given an equal mass corresponding to a relative unit of one. This study ignored accurate estimations of forces based on the figure's projected size, focusing instead on relative changes of physical parameters over time.

#### Recording and Rendering of Sequences

For each setting in the screenplay (gymnasium, stage and workshop bench) a visual scene was composed inside Unity. Textures, lighting and camera effects (e.g. field of view) were designed to appear very realistic in order to emphasise the physicality of the figure. Props in the final scene (mug, ruler, pencil, lamp, paintbrush) serve to point out the size of the figure (i.e. roughly as tall as a mug). The animation was unable to run in real-time when all light sources and effects enabled; especially not at a high enough resolution to export to a video file. Therefore each sequence in the screenplay had to be piloted at a lower quality and resolution setting. Controller

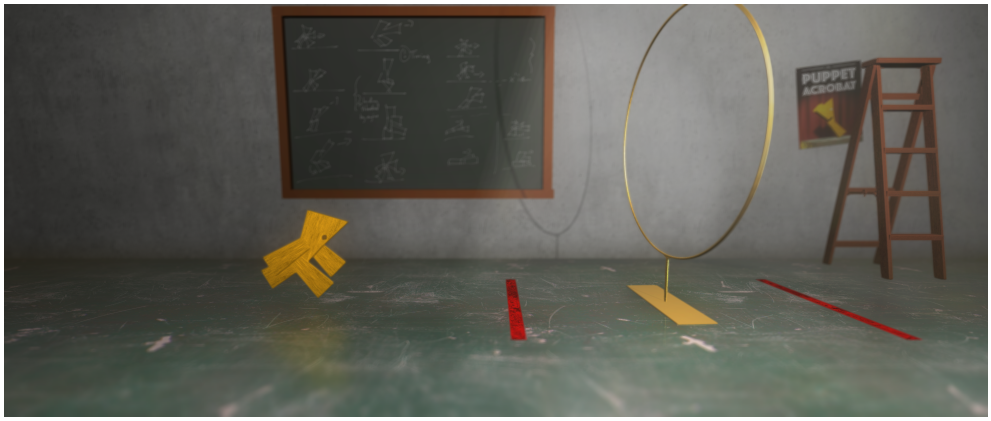


Figure 5.6: A fully-rendered frame from the 'gymnasium' scene.

input data at each frame was stored to a file and played back at lower speed. An uncompressed 4k resolution bitmap was stored along with physical data for each frame in the recorded sequence. Frames could then be converted into a high-resolution video and physics data was encoded as a multi-channel PCM wave file at 60Hz. Figure 5.6 shows a still from the final film.

### 5.3.4 Screenplay

A short screenplay was written by the author for an approximately two-minute film titled 'Nailex's big break' based on the animation system described above. The wooden figure was given a name ('Nailex'[4]) and described as an acrobat that is due to perform at an important show. The screenplay consists of five scenes based on three settings: a gymnasium in which the character practices jumping and performing somersaults through a large upright ring, a stage that contains a similar ring, on fire, suspended over a deep chasm, and a workbench that the figure falls down onto in the last scene. The protagonist ultimately fails at performing a single jump despite a more-or-less successful practice scene, following a contrasting dream sequence in which the character successfully completes a series of elegant jumps through the burning ring. Nailex is thus given anthropomorphic qualities with prescribed emotions (i.e. confidence, shock, fear, disappointment). A description of each scene is provided in Table 5.4 and the screenplay is appended in Appendix C.

Sound effect cues of 'applause' and 'lights switching on' have been added to the screenplay to emphasise that Foley performed for the character should not be designed to be the only sources of sound in the scene, despite these never being added to the soundtrack. This is to help detract attention from the fabricated nature of the

[4]Named by Foley artist John Roesch upon seeing a draft version of the animation.

| Scene # | Setting | Duration | Description |
|---|---|---|---|
| 1 | Gymnasium | 0'41" | Practicing acrobatics routine, running and jumping through ring multiple times |
| 2 | Stage | 0'34" | Strutting to centre of stage, gets a fright upon seeing burning ring, runs back to side of stage |
| 3 | Stage | 0'20" | Dream sequence performing elegant somersaults through ring multiple times |
| 4 | Stage | 0'16" | Striding towards centre of stage, stumbling, failing jump and falling down gap |
| 5 | Workbench | 0'15" | Falling from above onto centre of workbench, recovering, limping slightly and shuffling away from centre |

Table 5.4: Overview of scenes as featured in the screenplay for the animation.

experimental environment and approach the task as a regular Foley performance scenario.

### Event Types

While having a clear narrative thread the screenplay was devised to include a variety of contrasting movements that would make interesting case studies in the later analysis of performed soundtracks. These could be categorised as *repeated events*, *elongated sequences* and *unique events* at key narrative points. Repeated events primarily pertain to jumping actions and their individual components: preparing to jump, in-air movements (e.g. somersaults) and landing. Elongated sequences refer to sections of locomotion that span over an intentionally long duration. Each such sequence is based on a contrasting style of walking as directed in the screenplay: 'strutting', 'striding' and 'shuffling'. Unique events are moments of the figure's movement that occur in isolation at key points in the narrative. These include head and body movements as the character scans the stage and notices the burning ring, freely rotating limbs while falling down the middle of the stage, and the recovery after falling onto the workbench.

While the screenplay gives high-level directions for the movement of the character, these have been interpreted freely by the author in the piloting of the figure in order to achieve natural and evocative motion. Therefore other small movements within the interpretation of each event might have an effect on a participant's interpretation on the scene. This is in addition to unique events described above, which have intentionally been emphasised (or *spotted*) in the screenplay.

### 5.3.5 Computational Sound Models

The sound models have been kept intentionally simple, with each model being driven by a single variable parameter. Unlike the creaking door model presented in the previous chapter, these are all designed to react to physical parameters rather than subjective features. The purpose of this is to keep the dimensionality of the sound models low so as to constrain the focus of the study to the gestures controlling the models in response to depicted physical movement, rather than more complex timbral manipulation. Some fixed parameters, described below, were exposed to the participant to customise some aspects of the sound models. Participants could alter these at any point in the study, but were unable to dynamically vary their settings.

#### Resonators

A basic source-filter model is used as a basis for all the sound models. A set of five resonators (one for each limb of the animated figure) which are used as a filter for the corresponding impact and scrape sources. The friction model generating squeaks is passed through all of the resonators in parallel in order to maintain consistency with the rest of the sound sources. See Figure 5.7 for an illustration.

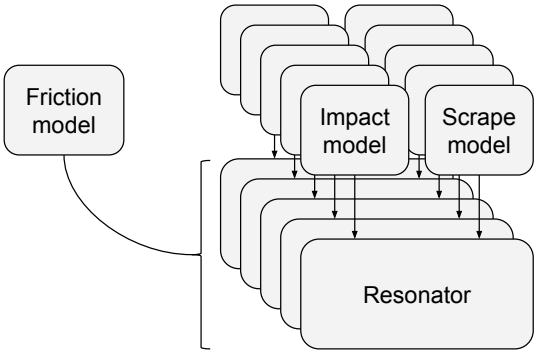


Figure 5.7: High-level source-filter architecture of sound models.

The resonators are banded waveguides Essl et al. (2004) each consisting of three bands. The frequency ratios of the bands are static[5] but are scaled based on a fixed *size* parameter. A *resonance* parameter controls the Q values of the band-pass filter for each band. These two fixed parameters can be set individually for each set of resonators. As this study doesn't take into account mass and density properties of the rigid bodies these parameters give the participants the opportunity to customize

[5]based on the modal frequency ratios for a uniform wooden bar accessible at http://csounds.com/manual/html/MiscModalFreq.html

an aspect of the soundtrack to their liking without it having an impact on the recorded performance data.

### Impacts

The impact model consists of a noise source that is passed through an exponential envelope with a duration of approximately 4 milliseconds (which is the length of the interval between successive frames, i.e. the rate of incoming physics data). The envelope is scaled by the incoming force parameter. As the force parameter is only sent on the entrance of a collision each new value simultaneously triggers the envelope after setting the scaling factor.

### Scrapes

The scraping source is based on a basic implementation of the model outlined by Van Den Doel et al. (2001). Pink noise is passed through a bandpass filter with a variable Q-factor and centre frequency. Frequency and Q-factor both increase as a factor of the velocity parameter that is passed into the model. Velocity values are interpolated to match the incoming rate of 60Hz. Usually a second parameter of continuous *normal force* affects the sound of the scraping, however this was omitted to keep the control 1-dimensional, thereby simplifying the performance data and keeping consistency with the physics data obtained from the game engine.

### Squeaks

The squeak source is based on the stick-slip friction signal chain described in Chapter 4. However, rather than having a timbre-based perceptual abstraction the model is based on a basic behavioural layer where velocity is mapped to the pitch of the impulse train. Like in the scraping model, this model is driven by a single parameter (*angular velocity*) and does not take into account effects of continuous normal force, in order to constrain the dimensionality of the collected data. Thus, to increase the believability of the model a simplified version of the *physically-inspired control layer* (see Section 4.4.3) was implemented (N.B. as a *behavioural abstraction* this time), where the crossing of a fixed velocity threshold causes the pitch of the impulse train to jump upwards, emulating the sound of mode lock-in Serafin (2004).

Some additional fixed parameters were implemented to allow the participant to customize the friction sound. *Stiction* corresponds to a variable velocity threshold, beneath which no sound is produced by the impulse generator. *Brightness* is implemented here by simply controlling a low-pass filter on the output of the impulse generator (rather than affecting the pulse shape). Finally, an *intensity* parameter scales the incoming velocity, affecting the overall pitch of the resulting squeak (lower

values corresponding to a 'creakier' sound and higher values more likely to produce pitches in the upper register of the mode lock-in emulation).

### 5.3.6 Enactive Hardware Interface

The core objective behind the design of the physical performance interface was for it to be approachable by the target group of people with professional Foley experience. This was deemed particularly salient here as the primary object of interest was not the capability of the interface, but rather the content and structure of the resulting performances, and how these related to the moving image.

Testimonies from the questionnaire indicated that previously developed metrics of *nuance*, *range* and *repeatability* applied to the target group: a high range of timbral diversity with a low entry-fee for nuanced control were deemed to be important factors.

The design of the hardware interface for controlling the models was based on an *enactive* approach, whereby incoming physical parameters were mapped directly to corresponding parameters of the sound models Essl and O'Modhrain (2006). The purpose of this was to keep the performed sensor data as close as possible to the physical data extracted from the animation system. The first prototype of the interface was replaced by a second iteration due to limitations found in the sensor bandwidth and further design considerations informed by responses to the questionnaire with Foley artists.

#### First Prototype

Both prototypes consisted of two fundamental components: a set of five force-sensitive and touch-capacitive pads to drive the impact and scrape models, and a crank-like interface resembling a wooden corkscrew for driving the squeak model. In the first prototype five Touchkey sensors (McPherson, 2012) where cut to a rectangular area of approximately 30 by 60 millimeters and arranged such that two fingers on each hand could play a set of two vertically orientated sensors while either thumb could play a horizontally placed sensor. The touch-capacitive components were used to detect velocity of each finger moving along the surface, driving the corresponding scraping model. A force-sensitive resistor was placed underneath each sensor to detect a pressure value, which was passed to the impact model on each new touch detected on the corresponding surface. While this produced acceptable results the pads were too small to generate all but very short scraping gestures and the fixed layout had a limiting effect despite being based on ergonomic incentives.

The crank was based on a rotary encoder with 90 slits. The upper part of a wooden corkscrew was repurposed as a handle that slotted into the rotary encoder. Velocity was derived from the readings of the rotary encoder and used to drive the

squeak model. The discrete stepping inherent to the rotary encoder had a highly degrading effect on the sensitivity of the interface, making it difficult or impossible to produce nuanced trajectories. Furthermore, the way in which the wooden handle was connected to the encoder (a narrow metal shaft) caused the interface itself to produce undesired sounds that were difficult to mute due to the physical structure of the rotary encoder.

Sensor processing was programmed on an Arduino Uno microcontroller, which sent data via serial to a host machine, where each sound model was implemented and running in Puredata. Serial communication resulted in noticeable latencies between action and sound and occasional dropouts.

Figure 5.8: The *crank* interface controlling the squeak model.

### Final Prototype

The second iteration is based on the same principles of touch and force-sensitive pads and a rotational interface, but several changes were made to mitigate the limitations described above. The microcontroller was replaced with the Bela platform (McPherson and Zappi, 2015) (Moro et al., 2016), which offered hard real-time audio and sensor processing at much higher resolutions and lower latencies than was achievable in the previous configuration.

The rotary encoder was replaced with a potentiometer, which in combination with the higher resolution provided a much more reliable velocity reading (albeit at the cost of finite rotations). This further allowed the interface to resemble an entirely wooden contraption by hiding all internal components (see Figure 5.8), while simultaneously eliminating mechanical noise caused by the previous prototype.

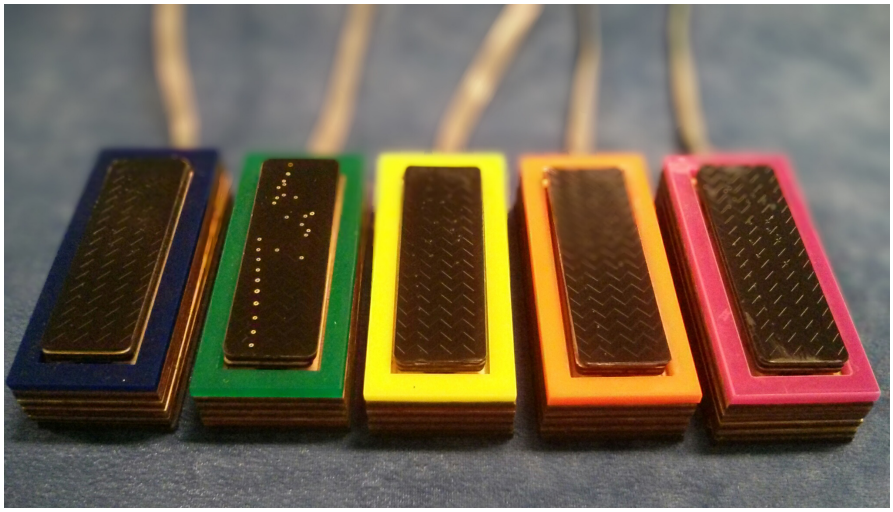

Figure 5.9: Touch and force-sensitive *pads* controlling the impulse and scrape models.

The pads were based on larger versions of the Touchkey sensors and used two force sensitive resistors in parallel (mounted below the capacitive sensor) to detect similar responses across the whole surface. In addition to this, the pads were mounted on individual colour-coded units, allowing them to be rearranged or picked up with ease (see Figure 5.9).

Running the sound model on the embedded hardware enabled low action-sound latency (less than 5 milliseconds) while preventing any jitter or dropouts in the sensor readings. Gestures corresponding to impacts and scrapes were measured at a sampling rate of 200Hz while the potentiometer was sampled at 22Khz.

### Mappings

As the interfaces were following an enactive design strategy, mediation between input dimensions and sound model parameters was minimised. Angular velocity measured from the potentiometer was used to drive the squeak model in units of degrees per second. Force readings and velocity measures were scaled to the ranges of the physical animation and directly controlling corresponding *force* and *velocity* parameters in the impact and scrape models. Two parameters were exposed in order to mitigate sensor limitations detected in informal trials. The FSR reading was raised to a settable exponent in order to circumvent the non-linearity of the sensor. Extra smoothing (also exposed as a settable parameter) was applied to the velocity reading in order to mitigate noise caused by differentiating at a low sampling rate and stepping artifacts caused by the spatial layout of the touch-capacitive sensors.

### 5.3.7 Synchronisation Workflow and Data Collection

Sensor data from Bela was transmitted via UDP (User Datagram Protocol) to a Digital Audio Workstation (REAPER[6]) where a VST plug-in transformed the incoming data into a 12-channel PCM audio signal at a sampling rate of 60Hz. The sensor data was downsampled on transmission in order to correspond to the graphical frame-rate and physics update rate of 60Hz (which was also the highest obtainable sample rate of the reference data). The sensor data was timestamped and logged locally to ensure accuracy of the transmission.

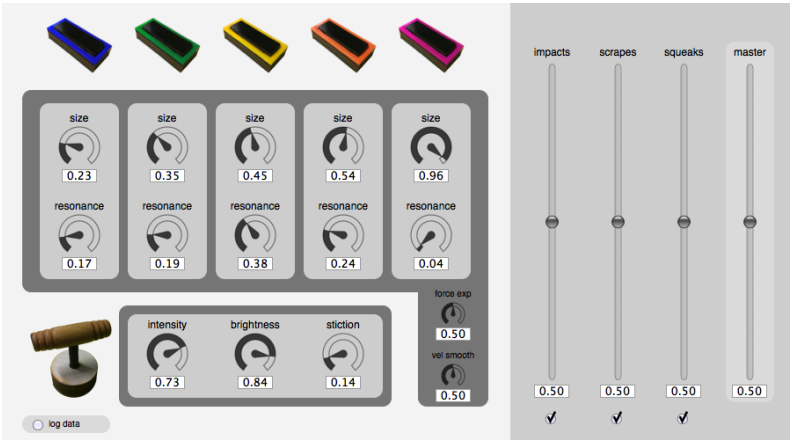


Figure 5.10: A screenshot of the plug-in user interface used for locally and remotely configuring the sound models and interfaces.

The original sampling rates were maintained for real-time performance in order to avoid degrading effects on the interaction. The sound model ran in parallel as a programme on Bela and as a VST plug-in on the host machine, allowing the participant to receive audio feedback with minimal latency.This facilitated high-quality real-time interaction while preserving conventional editing tools that the participant group was likely to be familiar with. Fixed parameter settings were shared across both instances and controlled via the user interface of the plug-in (see Figure 5.10).

REAPER facilitates the playback of high-resolution video as a concurrent track in the project. Jitter and drift across sound and image was tested using a light dependent resistor and an oscilloscope. Alternating sequences of white and black frames accompanied by impulses at each transition were measured while being played synchronously in REAPER. No observable jitter or drift was observed, however frames would only be partially drawn if computational resources were strained.

[6]http://www.reaper.fm

## 5.4 Summary

The experimental environment described in this chapter was designed with the aim of collecting a rich set of data suitable for a detailed investigation. The next chapter presents the findings of this study, discussing their implications on future integration technologies of computational audio for interactive environments.

# Chapter 6

# Analysis of Performed Soundtracks to a Procedural Animation

This chapter presents the procedure and analysis of a practical study based on the experimental environment presented in the previous chapter. Six participants with extensive experience in performing Foley were recruited to perform soundtracks for a two-minute animation. The physics-based nature of the animation made it possible to analyse performed soundtracks in light of a common source-model integration strategy (wherein physical data extracted from the animation are directly mapped to corresponding parameters in the sound models).

The primary aim of the analysis presented in this chapter is to obtain a general understanding of how intentional stylisations diverge from the physical data that would typically be used to integrate such models. Findings of this analysis can inform the development of future integration strategies that take performed stylisations into account, if and when they occur.

Section 6.1 provides a technical overview of the configuration and procedure of the study. Section 6.2 summarises general observations made in regard to the animation, sound models and physical interfaces featured in the study. Section 6.3 provides an overview of how soundtracks were structured by the participants, with a particular focus on their categorisation of sounds across separately performed *takes* or *layers*. Section 6.4 explores some of the emerging patterns and themes observed across participants' soundtracks. Section 6.5 proposes potential integration strategies based on five brief case studies from the performed soundtracks. Finally, some conclusions are drawn in Section 6.6, which are discussed in more detail in Chapter 7.

| | Years exp. | Num. Projects | Professional Role |
|---|---|---|---|
| P1 | 5-10 | 10-20 | Foley Artist |
| P2 | 10+ | 20+ | Sound Designer |
| P3 | 5-10 | 20+ | Foley Editor |
| P4 | 5-10 | 20+ | Foley Mixer (prev. Foley Artist) |
| P5 | 10+ | 20+ | Foley Artist |
| P6 | 10+ | 10-20 | Sound Artist (prev. Foley Artist) |

Table 6.1: Experience and Professional Roles of Participants

## 6.1 Configuration and Procedure

### 6.1.1 Participants

Six participants were recruited to take part in this study. Each participant had more than five years of professional experience and worked on more than 10 projects ranging across feature-length films, short films, theatre, television, games and art installations. Every participant had experience performing Foley to animation. Two participants had experience playing musical instruments. Participants were between 25 and 60 years of age, two participants were female and four were male. Table 6.1 shows the years of experience, numbers of projects worked on and professional roles for each of the participants.

### 6.1.2 Physical Setup

The study took place in a sound-proof studio (the control room in Queen Mary University of London's *Media and Arts Technology Studios* (Morrell et al., 2011)). The research investigator was present at all times throughout the study, sitting next to the participant and operating the Digital Audio Workstation (DAW). The participant sat in the centre of the studio desk in front of a screen that was used to display the animation. The physical interface was placed on the desk below the screen and was concealed for the first part of the study (the participant was allowed to rearrange the layout of the interface's components throughout the remaining parts of the study). Two further screens were placed to the right of the desk. The furthermost screen displayed the DAW's time-line which was exclusively operated by the investigator. The other screen displayed a user interface for controlling the fixed parameters of the sound models, which could be controlled by both the investigator and the participant. A video camera was set up in the corner of the room, facing the participant and the physical interface. Video and audio footage was captured for the duration of the study.

### 6.1.3 Procedure

The study consisted of three parts and took between one-and-a-half and two-and-a-half hours to complete. The purpose of the first part was to familiarise participants with the animation, the sound models and the physical interface. This was followed by the core practical part of the study, which involved iteratively performing soundtracks to each scene of the animation using the physical interfaces provided. Finally, in the third part of the study participants were shown the animation three times with different soundtracks.

The following subsections provide a concise overview of each part of the experimental procedure.

**Part 1**

1. Participant is shown the screenplay (see Appendix C)
2. Participant is shown the animation (with no sound)
3. Research investigator asks the participant open questions about how they would normally approach performing Foley for the animation:

   *How would you approach the Foley soundtrack for the wooden character?*

   *What props could you imagine using for producing these sounds?*

   *What challenges can you see arising in performing Foley for this character?*

   *Have you worked on similar scenarios before?*

4. Participant is asked to listen to a set of sound examples representing the range of the sound models. All of the example sounds were performed on the corresponding interfaces and rendered as a fixed waveform. The participant is encouraged to comment on the sounds, particularly their suitability for use in conjunction with the animation.
5. Participant is presented with the interface and asked to explore it without any further instructions.
6. After a few minutes of exploration the research investigator asks the participant a set of open questions:

   *Does the interface feel natural to play?*

   *Does the correspondence between your actions and the resulting sounds make sense to you?*

   *Do you feel that you have a good sense of control over the sounds?*

7. Participant is given a detailed description of the physical interface.

8. Participant is shown the software interface (see Figure 5.10 in the previous chapter) and given a detailed description of each parameter.

9. Participant is asked to tweak the settings for each component of the physical interface (with the research investigator's assistance if needed), to suit the sounds they intend on producing for the soundtrack.

10. Participant is given the opportunity to further explore and tweak the interface until they are satisfied with the settings

**Part 2**

1. Participant watches the animation one more time before proceeding to perform the soundtrack on a scene-by-scene basis.

2. For each scene the participant is allowed to record as many takes as they want, and to layer multiple impact and scraping sounds if they desire to do so. Only one instance of the 'squeak' model is available (to maintain consistency with the physical data extracted from the game engine), thus precluding the ability to layer multiple performances polyphonically. A 'count-in' of four sinusoidal impulses is given before each take to anticipate the start of each scene.

**Part 3**

1. Participant is shown the animation three times: twice with soundtracks by previous participants and once with the physical reference soundtrack (in randomised order). The first participant was shown only two soundtracks, one of them having been performed by the participant of a successful pilot study (also a professional Foley artist).

2. Participant is asked to rank each soundtrack in order of preference.

3. Participant is asked to comment on aspects of each performance that they liked or didn't like

### 6.1.4 Role of the Investigator

The investigator fulfilled a number of roles throughout the study including:

- Operating the DAW while the participant completed the primary task
- Leading the participant through all stages of the study procedure
- Fostering an environment for open discussion while completing the tasks

The investigator took on some roles that would conventionally be taken by the Foley editor and Foley mixer. These involved constant coordination with the participant in order to operate the workstation without interrupting the task. Because sensor data was logged as PCM waveforms, it was possible to implement conventional DAW workflows, including the recording of multiple takes, muting and unmuting tracks, setting markers, and so forth. Care was taken to name tracks according to terms suggested by the participant (e.g. 'footsteps', 'whooshes').

At key moments the investigator would encourage the participant to comment on certain choices made while avoiding biasing factors in the collected data. These included occasions were participants discarded a take, and upon completion of the soundtrack for each scene.

### 6.1.5 Collected Data

Sensor data was logged throughout the whole study. In addition, sensor data for each take was recorded in the DAW as 16-bit multi-channel waveforms sampled at 60Hz. These could later be compared directly to the physical reference data on a per-scene and per-source basis.

Despite the investigator operating the DAW, the structure and meta-data of the resulting workstation projects provided insights into the way that the participants approached the soundtrack, for example how individual takes were labeled or referred to and the order in which they were recorded.

As participants were encouraged to make comments throughout the practical stage of the study, audio-visual footage of the study provided some key insights pertaining to specific elements of each soundtrack. This complemented a more technical analysis of the sensor data.

### 6.1.6 Approach to Analysis

The collected data was analysed in three stages. The aim of the first stage, presented in Section 6.3, was to gather a general overview of how each participant's soundtrack was structured. This included exploring the *vertical* structure of the soundtrack as it was recorded; in other words, how each take was labeled, which sensor configurations were applied and which sound models were used. In addition to this, general observations about the temporal structure were made in light of the physical reference data.

The second stage involved identifying emerging patterns across the corpus of performed soundtracks. Key observations made in regards to the temporal structures, sound-source and action-source relationships, and their dependence on various information from the animation (i.e. physical reference data, animation state and narrative

context) are presented in Section 6.4.

Finally, a set of case studies were chosen in order to get an understanding of the technical requirements for automating the corresponding behaviour (i.e. integration strategies) in an interactive environment such as a game engine. These are presented in Section 6.5.

The next section provides a brief overview of general observations made regarding the general experimental setup, including the sound quality of the models, the physical interface, the animation and the general workflow of the synchronisation task.

### 6.1.7 Supplementary Audio-Visual Material

Videos of the animation with each of the participants' soundtracks and the physical reference soundtrack can be found in the on-line supplementary audio-visual materials (see Appendix E.3). The videos contain overlays of frame numbers relative to each scene. This corresponds to the frame numbering used in the diagrams below.

## 6.2 General Observations and Perceived Limitations of the Experimental Environment

### 6.2.1 Animation

While all participants had worked on similar types of animation, each participant commented on the complexity of the animation and its implications on performing Foley.

Four participants (P1, P2, P5, P6) contrasted the character's movement to (filmed) human movement, where locomotion would normally be more predictable and rhythmic. P5 mentioned that 'when you are doing [Foley for] animation ... there is no warning for what they do ... when Foleying a human you can follow the shoulders and you know what is coming. But [in this animation] all of a sudden he just stops or he just skids', marking unpredictability as a quality common to the medium of animation in general. The other two participants suggested that the animation possessed both anthropomorphic and animal-like qualities, therefore making it particularly challenging or unusual to perform Foley for. P1 related the movement to that of an animal or a dog, stating 'I wouldn't walk him as I would walk a human. But his pacing is interesting'. While for human movement this participant would normally try to perform each footfall, the fast movement of this character combined with its conversely slow forward velocity meant that if they 'walked it literally like every foot down it would sound wrong' suggesting that they would instead 'go for rhythm and feel' in the performance.

Some participants perceived the character's movement as being mechanical in nature. P4 suggested that this would result in the soundtrack carrying more importance in conveying emotional states from the screenplay: '... emotional elements like fear, disappointment, he hurts himself et cetera. I would try to bring those out, try to add to that. Given that he's quite a rigid structure in animation, I would want to do as much as possible in order to sort of really give that to the audience.' P6 contextualised this in specific character motions, such as cases where the character falls over and slides: 'the difficult bits would be the trips and slides. there's a kind of rhythm to it that's ... non-human ... it's much more discontinuous and much more broken et cetera ... although there's a kind of humanness to that character, there's something a little bit more mechanical.'

### 6.2.2 Layers

When asked about how they would normally approach the Foley for this animation, each participant described the desirable soundtrack in terms of *layers* or *takes*. As is evident from the above statements, there was a clear focus on the character's locomotion and approaches to performing a corresponding layer of *footsteps*. Due to the unusual (or non-human) nature of the locomotion some participants mentioned that they would not use their feet to perform these sounds (as would normally be done when performing human footfall). Participants suggested using wooden blocks (P1,P2,P4,P5) or clothes pegs (P3). Three participants (P1,P4,P5) referred to a 'loose' quality in the character's movement, and therefore suggested using a 'weird wooden clackity' prop (P1) or a 'wooden puppet' (P4,P5) to capture this in the soundtrack. P4 elaborated further on the qualities of the prop, stating that 'the thing that would be quite important to me would be something with some inbuilt movement'. While P5 suggested using such a prop as a secondary layer to compliment a more detailed rendering of the character's footfall, P1 and P4 ascribed this to a primary role in capturing the character's movement. This is evidenced in the following statement by P4: 'I think in something like this a very important one [layer] would be something covering his movement - just his general movement, not his footsteps, not his arms hitting things'.

Three participants suggested strategies for capturing the rotational movement of the character's limbs as it walks or performs somersaults in the air. P3 suggested the use of 'squeaky' sounds, while P1 and P5 suggested the use of 'wood on wood' slides to capture this movement.

Four participants (P1,P3,P4,P6) distinguished between the aforementioned movements and more singular actions such as the character coming to a halt and sliding, fast rotations in the air and its head colliding against surfaces. P4 referred to the corresponding elements of the soundtrack as 'spot effects' - a common term in Foley

and audio post-production that is used to refer to distinct sounds in the scene (as opposed to recurring or continuous sounds) (Ament, 2009).

While the required number of layers suggested by participants ranged between three and eight, two participants (P3,P4) emphasised that the character's soundtrack should be kept 'concise' and not be 'annoying' for the audience.

### 6.2.3 Sound Models

Having watched the animation and thought about how they would approach the soundtrack, the sound examples did not always align with the participant's expectations. Three participants (P1,P2,P5) suggested an element of serendipity in their normal working process, where props would be selected on a trial-and-error basis until finding one that produces a desirable result when matched to the moving image. P1 referred to the 'prop cupboard' in a typical Foley studio, where a multitude of unique objects are at the artist's disposal. P2 noted that 'every time you're doing it [finding an appropriate prop] it's an individual problem'. This reflects observations from the survey with Foley artists (presented in Section 5.2), wherein most respondents claimed that the ideal sound qualities of a prop depend largely on the context of the visual counterpart in the image. Therefore, it felt unnatural for some of the participants to work with a prescribed and constrained palette of sounds. Nonetheless, all participants thought that the types of sounds presented to them were suitable for the animation, albeit different from how they would have originally approached the soundtrack.

Three participants (P1,P4,P6) suggested that the quality of the sounds would have an effect on the consequent perceived context of the animation. P1 elaborated on this by referring specifically to the target audience of the animation (which was intentionally left ambiguous in the screenplay): 'I would often pick sounds depending upon who this is aimed for. These sound more playful ... so if it's a children's afternoon TV programme then the [sounds] would be quite appropriate'. P4 remarked on the 'simplicity' and 'musicality' of the sounds, noting that they were 'closer to the role that music might perform typically in an animation', where 'every action is expressed with music'. P6 referred specifically to the sounds produced by the squeaking model as adding a 'slightly comic element', noting that this would normally be the result of an 'aesthetic' or 'stylistic choice' for the animation (imposed by a director, producer or supervising sound editor).

Three participants (P1,P2,P3) associated a 'vocal quality' with the sounds produced by the squeaking model, with P2 noting that they were 'like somebody talking'. P2 and P3 remarked that these would add an element of 'characterisation'. P1, whose main profession was a Foley artist, related these sounds to a vocal 'reaction' by the character, but noted that they would typically refer to this as a 'sound effect' and not

'something I would do'.

The output of the sound models was perceived by two participants as being simplistic (P4) or synthesised (P6). Three participants (P1,P5,P6) associated the output of the scrape model with the sound of 'breath', particularly at high resonance settings ('I didn't like the scrapes very much because they sound a bit like somebody is blowing across a glass'). In the practical stage of the study this model was commonly used to recreate a variety of sources including wind 'whooshing' and wood sliding against wood. This can maybe be attributed to the simplicity and, therefore, ambiguity of the sound model.

### 6.2.4 Interfaces

#### Pads

Referring to the pads, P1 stated: 'I'm used to picking something up rather than performing on an item - it feels very unusual ... it feels like that's doing the work, rather than me.' P6 remarked specifically on the responsiveness of the sensors, noting how it didn't reflect physical relationships from the everyday environment: 'with a smaller amount of change of impact I get a larger amount of change in amplitude.' On one hand this reflects a technical limitation of the sensors used in this interface (the force-sensitive resistors do not have a sufficiently large range and have a non-linear response). On the other hand, the general interaction principle was considered unnatural for the sounds that the interface produced. One of the participants demonstrated how they would have expected to use these sensors to produce impact and scraping sounds by picking them up and knocking them against the table with their sides (see Figure 6.1). Instead, the interaction afforded by the pads was likened to more conventional MIDI interfaces such as musical keyboards by two of the participants (P2,P6).

The parameterisation and real-time interaction with the collision model was also deemed by some participants to lack in nuance, resulting in being perceived as more music-oriented than timbre-oriented. P5 noted how using two pads with varying size and resonance parameters caused the character to sound 'like he's got a pegleg', while using a single sensor resulted in each leg sounding too similar. As a mitigation strategy the participant decided to use two pads with slightly different size and resonance parameters. P6 noted: 'I was being forced to make more musical choices than timbral choices - choices about pitch rather than the quality of the sounds.' These observations suggest a lack of range and nuance of the interface, perhaps caused by a perceived bias towards pitch in the model's parameter space and the inability to performatively vary the sound of each collision beyond its intensity in volume.

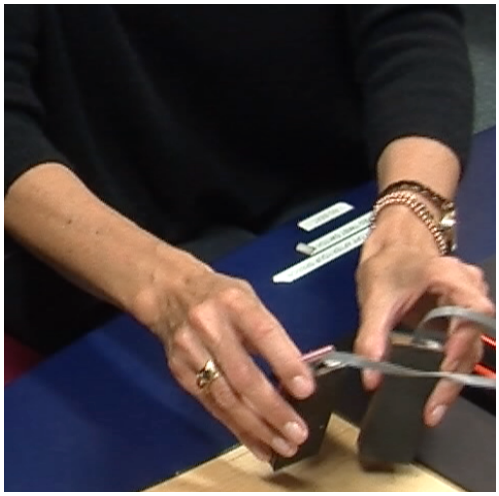

Figure 6.1: Demonstration by one of the participants of how they would have expected to produce collision and scraping sounds.

#### Crank

In contrast, the crank received an overwhelmingly positive response from the participants; four participants said that they preferred it over the pads while the remaining two did not state any preference. Two participants (P2,P5) expressed their desire to own such an interface to produce squeaking sounds, one of them (P2) noting that it 'has so many applications'. In a comparison to the pads, P6 marked a strong contrast between musical and Foley-like qualities: 'It felt more like trying to get the sound out of a concrete material. This is Foley, that's very natural.' From a technical point of view, this is interesting as the interface has a very low dimensionality and, in terms of hardware, uses a very simple sensor (a potentiometer) to detect movement. This suggests that the tangibility of the interface and/or its resemblance to a real corresponding source (a wooden crank or corkscrew) had a strong effect on the perceived quality of the interface.

### 6.2.5 Temporal Detail

While all participants were comfortable with the task of performing all of the Foley with little or no subsequent editing, all participants noted that they would normally rely on an editor to correct the synchronisation. 'Good' or 'perfect' synchronisation was regarded by many as a trait of a highly experienced or 'old school' Foley artist. Three participants (P1,P2,P3) noted that with the emergent dependency on digital tools it was less important to achieve perfect (i.e. frame-accurate) synchronisation in the performance, this task instead being left to the Foley editor.

Table 6.2: Rankings of human-performed (HP) and physics-driven (PD) soundtracks by participants.

| Rank | P1 | P2 | P3 | P4 | P5 | P6 |
|---|---|---|---|---|---|---|
| **1** | *HP (pilot)* | *PD* | *HP (P1)* | *HP (P3)* | *PD* | *PD* |
| **2** | *PD* | *HP (P1)* | *PD* | *PD* | *HP (P3)* | *HP (P5)* |
| **3** | | *HP (pilot)* | *HP (P2)* | *HP (P2)* | *HP (P4)* | *HP (P4)* |

Indeed, following the practical stage of the study most participants commented that their performances were not ideal in their current state due to lack of editing and mixing that would normally be carried out after recording all the Foley tracks. This makes it difficult or impossible to draw any conclusions about intentional latency or anticipation on the microscopic level of discrete event-sound relationships (e.g. whether an impact was performed a few frames before or after its corresponding event for aesthetic reasons). It also raises the question of whether such effects (should they be a commonly occurring phenomenon) are the result of iterative editing by Foley editors rather than temporal nuances in the real-time performance of sound to the moving image.

### 6.2.6 Listening Evaluation

Five out of six participants (P1,P3,P4,P5,P6) correctly identified the computer-generated soundtrack.

Participants were asked to rank each soundtrack according to their preference. Rankings are shown in Table 6.2.

Every participant highlighted differences between the physics-driven soundtrack and a human performed one, before being told that one of them was not performed by a human.

Every participant commented on the accuracy of the audio-visual synchronisation in the physics-driven soundtrack (‘very remarkable sync’ (P1), ‘seemed the most accurate’ (P2), ‘the sync was great in this ... if I had to guess I would say that this was done by a really experienced Foley artist’ (P3), ‘I only know a couple of people ... that would get the sync to work so well’ (P4), ‘the sync is remarkably precise’ (P5), ‘clearly has the best sync’ (P6)).

Two participants (P2,P5) associated a higher degree of realism with the physics-driven soundtrack. P5 found this to be a positive property that they associated with their own approach as a Foley artist stating ‘maybe I’m more of a realistic Foley artist, so I liked that element of realism and perfect sync’.

Three participants (P1, P3, P4) associated sterility or lack of character with the computer-generated soundtrack, before being told that one of the soundtracks was not performed by a human. P1 found the soundtrack ‘too sterile, despite the perfect

sync ... whereas the other one [human-performed soundtrack], this one I can relate to, I feel like it was made by one of us'. P3 noted that 'there's been a real attempt to make his walking work ... but as far as making him or her into a character ... there's maybe slightly less of that.' P4 questioned whether 'there was copy-and-pasting on the squeak track'. This points to a lack of variety in the physics-driven soundtrack, where repetitive sequences in the walking animation produce the same sounds, despite its physics-based nature.

P4 continues '... this led me to wonder: how does this person have so much control over the slides?', referring to moments where the character slides along the floor for extended periods of time. Indeed, every participant pointed out that they had difficulty performing these sequences due to the size of the pad interface, which precluded long, uninterrupted gestures. In the same vain, after being told about the machine-generated soundtrack, P3 commented on the impact sounds mentioning that 'the transients make more sense here', whereas the human-performed soundtracks tended to contain 'accidental pops' which they associated with difficulties in using the interface.

## 6.3 Structure of Performed Soundtracks on a Per Participant Basis

### 6.3.1 Overview

Each participant decided to layer multiple takes in order to construct the soundtrack to each scene of the animation, rather than playing all parts of the interface at once. Perhaps more surprising is the number of layers that were used for each scene, and the consistency with which they were categorised by the participants.

Unlike the physical reference, which was structured according to 'impacts', 'scrapes' and 'squeaks', participants used between four and six separately recorded tracks for each scene. These were labeled according to more specific categories of the character's movement, such as 'tumbles', 'footsteps' and 'body hits'.

Every participant dedicated a track that corresponded to the character's locomotion (labelled 'footsteps' or 'feet') in each scene, which were performed using one or more of the pad sensors. Similarly, each participant dedicated track to squeaks performed using the crank interface. Less common action-sound couplings included 'whooshing' sounds for mid-air jumping sequences (P3) and irregular, playful sequences of impacts as the character is falling (P4).

These classifications of the soundtrack are an interesting observation as they refer to *action-source* relationships rather than *source-sound* relationships. In other words, while the reference soundtrack was based on physical interactions between individ-

| Label | Sound Models / Sensors | Scenes |
|---|---|---|
| Footsteps | Impact, Scrape (four pads, up to two simultaneously) | 1,2,3,4,5 |
| Lands | Impact, Scrape (three pads, up to three simultaneously) | 1,3,4,5 |
| Messy | Impact, Scrape (five pads, up to four simultaneously) | 3,4 |
| Slides | Scrape (two pads, one at a time) | 2,3,4,5 |
| Squeaks | Crank | 1,2,3,4,5 |
| Air | Scrape (two pads simultaneously) | 1,3,4,5 |

Table 6.3: Track structure for Participant 1

ual objects (the character's limbs, surfaces and hinges or joints), participants based their performances on specific types of the character's actions (walking, falling over, rotating, and so on).

In addition to this 'vertical' composition, participants' soundtracks also diverged from the physical reference data in their temporal structure. For example, during sequences of repeated physical movement (with regular intervals and values) participants were often found to omit or exaggerate the rendition of specific movements in the animation.

### 6.3.2 Participant 1

All of the categories in Table 6.3 were recorded in the order that they are shown.

**Footsteps** consisted of impacts and scrapes performed on two pads simultaneously and corresponded to impacts and scrapes of the character's legs on the floor during walking or running sequences. Impacts and scrapes were only performed for steady sequences of locomotion excluding transitioning states that contain more irregular steps and/or have a narrative significance identified by the participant.

**Lands** were performed using a different combination of pads (with different size and resonance settings). As the label suggests, performed sequences corresponded exclusively to moments in which limbs impact a surface, sometimes extending to short sequences of recovery (standing up) after collapsing.

As in the previous category **Messy / Offs & Lands** are sequences of impacts and scrapes performed on multiple (up to four) pads simultaneously. Sound behaviours are more imprecise and granular ('messy') and correspond to jumping and landing motions ('offs' and 'lands') in the third scene, and the character falling down the hole at the end of the fourth scene.

**Slides** consist primarily of scrapes and occur during transient or irregular sections of locomotion, such as shuffling limbs when approaching a ledge in scene 2, jumps and lands in scene 3 and sequences of recovery. The participant described these sounds as corresponding to the limbs scraping ('sliding') against each other as opposed to individual limbs scraping along the surface.

| Label | Sound Models / Sensors | Scenes |
|---|---|---|
| Footsteps | Impact (two pads simultaneously) | 1,2,3,4,5 |
| Extra Footsteps | Impact (same as above) | 2,4 |
| Scrapes (Floor) | Scrape (five different pads, up to four simultaneously) | 1,2,5 |
| Scrapes (Body Movement) | Scrape (two different pads, one at a time) | 1,2,3,4,5 |
| Squeaks | Crank | 1,2,3,4,5 |
| Extra Squeaks | Crank | 5 |

Table 6.4: Track structure for Participant 2

**Squeaks** represent the only track that was performed using the crank interface. Performed sounds are usually very close to the rotational movement observable in the physical reference data but are only applied to key moments when the character is moving its head (i.e. looking), during rotations and at transient states of locomotion (see Figure 6.2).

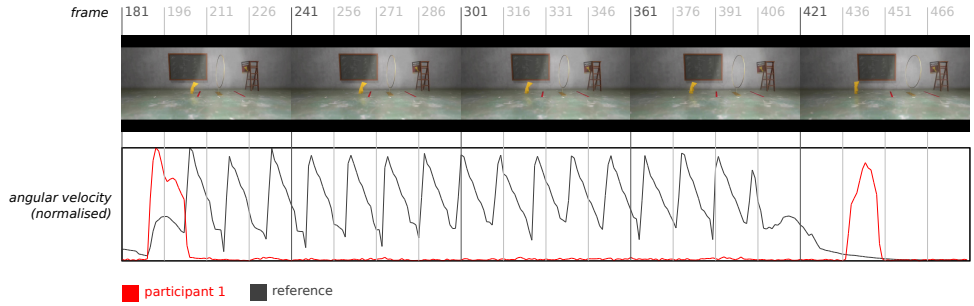


Figure 6.2: Squeaks performed by P1 mark transient states of the character's locomotion in Scene 1.

Similarly to the *Slides* track, **Air** corresponds to limbs scraping against each other. Performed sounds correspond exclusively to periods where the character is in-air (i.e. between jumping and landing).

### 6.3.3 Participant 2

**Footsteps** consisted of impact sounds of the limbs that generally corresponded closely to those generated by the reference track. Two pads were used with slightly different *size* parameters (resulting in two distinct pitches). Most sequences featured a repeated alternation between the two pads (A-B-A-B), regardless of the order in which limbs were impacting the surface. An auxiliary track labeled **Extra Footsteps** was used to separately sonify rapid leg movement at the beginning of the walking sequence in scene 2 and for a moment in the fourth scene where the character's head hits the wall during the falling sequence.

| Label | Sound Models / Sensors | Scenes |
|---|---|---|
| Impacts (Footsteps) | Impact, Scrape (single pad) | 1,2,3,4,5 |
| Impacts (Bonk) | Impact, Scrape (single pad) | 4,5 |
| Slides | Scrape (single pad) | 1 |
| Whoosh | Scrape (two pads, up to two simultaneously) | 1,2,3,4,5 |
| Squeaks | Crank | 1,2,3,4,5 |
| Vocals | Crank | 2,3 |

Table 6.5: Track structure for Participant 3

The participant differentiated between scrapes across the floor surface and rotational scrapes across two limbs. **Scrapes (Floor)** corresponded closely to surface scrapes extracted from the animation, but were omitted in the third and fourth scenes.

**Scrapes (Body Movement)** corresponded to the rotational movement of the limbs. A second track labeled 'legs' was recorded for the second scene, in which the participant wished to differentiate arm and leg movements.

**Squeaks** played a particularly important role in this soundtrack according to the participant, providing the soundtrack with an element of 'character'. On one hand they were 'tempted to give him a squeak ... every time he walks ... because that's going to add character ... if you give him an occasional squeak when he walks'. On the other hand the participant assimilated the sound of the squeak with a form of vocalisation for the animated figure ('this will give him character ... it's more like someone talking').

A further track of **Extra Squeaks** was used to emphasise key elements in the final scene of the animation. As only one squeak model was active in this study, summing two streams of sensor data together resulted in different and sometimes unexpected behaviour compared to the effect of layering to audio recordings of the output. The participant stated that they would have liked to be able to layer multiple separate squeaks, but was happy with the result achieved from summing two streams of sensor data.

### 6.3.4 Participant 3

Similarly to Participant 2, this participant used a single track labeled **Impacts (Footsteps)** to sonify all impacts between the limbs and the floor surface. While only a single pad was used to perform this track, this participant was one of two participants to consistently use two impacts to sonify each step of the character's locomotion. The reference data extracted from the animation also contains two impacts per step, whereby each leg makes contact with the ground (due to the symmetry of the movement) at intervals varying between zero (i.e. simultaneous occurrence) and one frame. Intervals in the performed walks were generally longer (up to ten frames) and had

greater degrees of variation (see Figure 6.3, which may either be a constraint of the interface or an attempt at exaggerating the slight asymmetricities in the movement as a stylistic choice. An additional track labeled **Impacts (Bonk)** was used to sonify the head impacting the wall at the end of the fourth scene, and the ground at the beginning of scene 5, using a different pad with a higher *size* parameter.

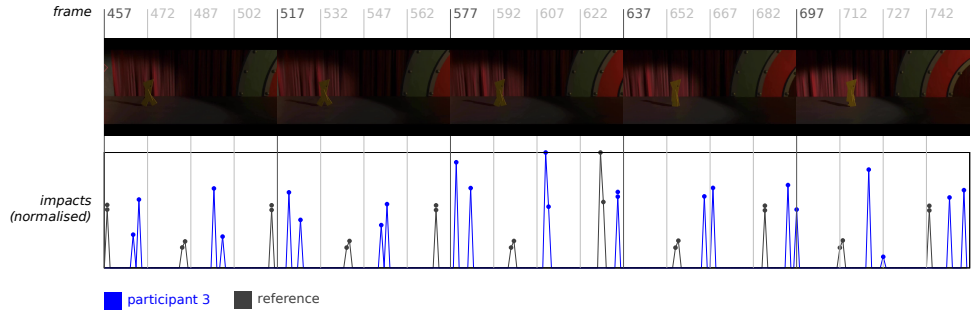


Figure 6.3: Dual impacts on each footfall performed by P3 correspond to physical reference data in Scene 2. Greater variability and range in intervals between impacts can be observed.

**Slides** were only present in the first scene and consisted of short scrapes to sonify longer sequences of the character sliding along the floor and rotations (changes in walking direction). The participant originally used a separate track to perform scrapes that were more or less consistent with the corresponding reference data, but decided to omit it at the end of the task due to the soundtrack sounding 'too full'.

The participant used the scrape model to render the sound of air **whooshing** as the character jumped through the ring in scene 3, and during the falling sequences in the fourth and fifth scenes. Two tracks were used in the final two scenes to blend between two pitches (*size* values) corresponding to the vertical position of the character while falling.

**Squeaks** were performed to all sequences of locomotion in a way that very accurately mirrored the corresponding physical reference data. This extended to some transient locomotive movements including jumping and landing, but did not follow mid-air movements such as somersaults. Instead, a separate track labeled *Vocals* was used to emphasise particular movements and events that the participant associated with vocal expression (e.g. physical exertion during jumps in scene 3 and shock in scene 2). It is also worth noting that this track was performed using larger hand movements resulting in them being identifiable by a higher range of pitch.

| Label | Sound Models / Sensors | Scenes |
|---|---|---|
| Footsteps | Impact, Scrape (four pads, up to two simultaneously) | 1,2,3,4,5 |
| Body hits | Impact, Scrape (four pads, up to two simultaneously) | 1,3,4,5 |
| Rotations | Scrape (two pads, one at a time) | 1,3,4,5 |
| Windy movement | Scrape (two pads, one at a time) | 1,2,3 |
| Squeaks | Crank | 1,2,3,4,5 |

Table 6.6: Track structure for Participant 4

### 6.3.5 Participant 4

**Footsteps** following each limb impact mirroring physics data, except for floor collisions upon termination of jumps, which were rendered separately in a track labeled **Body hits**. These were performed using a different set of pads (with different *size* and *resonance* settings). In addition to landing movements this track was used to perform an abstract granular sequence of impacts during the falling sequence at the end of Scene 4 (similar to P1).

The scraping model was performed separately from the impacts. The participant used the model to render moments of the character sliding along the floor (in the first scene) and mid-air rotations (corresponding to limbs sliding against each other). Up to three tracks were used simultaneously to record these sounds. A track labeled **Rotations** was used to perform the scraping model corresponding to limb rotations while the character was in the air. An additional track labeled **Windy movement** featured broader motions corresponding to general movement through the air with less detailed emphasis on quick rotational movement. During the falling sequence in Scene 4 the performance featured more erratic movement in an attempt to dramatise the character's emotional state as it fails to jump across the ledge.

**Squeaks** were performed in close correlation to the physical movement. Exceptions included some intentional stylizations including steadily increasing intensity as the character approaches the ledge in preparation to jump in Scene 4, and exaggerated movement during Scene 2 mirroring some of the performed behaviours other participants referred to as 'vocalisations'.

### 6.3.6 Participant 5

This participant's soundtrack bore the strongest resemblance to the physical data, both in the vertical structure (one layer for each sound model) and in the temporal correlation of physical movement.

**Footsteps** were always performed first using up to two pads simultaneously. Temporally, the performed impacts were remarkably consistent with the character's footfall.

**Squeaks** and **scrapes** also corresponded very closely to the physical data, with

| Label | Sound Models / Sensors | Scenes |
|---|---|---|
| Footsteps | Impact, Scrape (three pads, up to two simultaneously) | 1,2,3,4,5 |
| Scrapes | Impact, Scrape (four pads, up to four simultaneously) | 1,2,3,4,5 |
| Extras / Impacts | Impact, Scrape (five pads simultaneously) | 1 |
| Squeaks | Crank | 1,2,3,4,5 |
| Shooshing | Scrape (five pads simultaneously) | 4 |

Table 6.7: Track structure for Participant 5

| Label | Sound Models / Sensors | Scenes |
|---|---|---|
| Feet | Impact (two pads simultaneously) | 1,2,3,4,5 |
| Tumbles | Impact (three pads, up to two simultaneously) | 1,2,3,4,5 |
| Squeaks | Crank | 1,2,3,5 |
| Scrapes | Scrape (two pads, one at a time) | 1,2,5 |

Table 6.8: Track structure for Participant 6

the former generally corresponding to the combined angular rotation of the character's limbs and the latter to limbs sliding along the surface.

Only two additional layers were used, each for a single scene only. In these layers there was comparatively less correspondence to the physical data. **Extras / Impacts** consisted of granular bursts of impacts (using all five pads simultaneously) to emulate the sound of the character collapsing onto the floor after each jump. **Shooshing** corresponded to the rotational sliding of limbs against each other as the character falls down the stage at the end of Scene 4. Again, the movement is fairly broad and granular with all five pads used simultaneously.

It is worth noting that the majority of the soundtrack was based on first takes, with only two rejections throughout the whole synchronisation task. This participant was also the most experienced, having worked as a Foley artist for over twenty-five years.

### 6.3.7 Participant 6

The first layer recorded for each scene was labeled **Feet** and corresponded to the character's locomotion. Similarly to P3, this participant used two impacts per step, but on two pads with slightly different size parameters. An exception to this was in Scene 5 where the participant used only a single sensor to distinguish the character's shuffling motion.

A separate layer of impacts, labelled **Tumbles**, was used at transient moments of locomotion (starting, standing up and turning) and occasionally during mid-air movement. These gestures stand in contrast to those performed in the *Feet* layers in that they contain more arbitrary or granular series of impacts corresponding to singular events.

**Squeaks** were used in three scenes (1,2 and 5). They were generally applied while the character was walking or running, with exaggerated motions during transient states of jumping and landing. The sequence of the character running away from the ledge in Scene 2 featured exaggerated motions, corresponding to the state of 'fright' specified in the screenplay. Another interesting feature is a gradual fading of movement towards silence as the character walks towards the edge of the frame in the final scene.

**Scrapes** were used very selectively Scenes 1,2,3 and 5. Performed gestures corresponded to singular events such as jumping and landing, and occasional transient states of locomotion when the character was stopping, turning and sliding along the floor.

## 6.4 Emerging Patterns

### 6.4.1 Regular and Irregular Movement

Some interesting patterns can be observed when comparing the structures of the participants' soundtracks. Every participant started the synchronisation task for each scene with a 'footsteps' track, corresponding to sequences of the character's bipedal locomotion. While performed impacts generally corresponded to the character's footfall, all participants pointed out that they didn't deem their temporal synchronisation to be optimal and that they would normally rely on further editing to correct the timing of each impact (or set of impacts) to the corresponding footfall. Figure 6.4 illustrates a sequence of regular locomotion in which performed timings varied by up to 18 frames (300ms) across all participants. It is also worth noting that the vast majority performed impacts in this particular sequences were performed later than the corresponding collisions in the physical reference data.

In addition to footsteps, each participant also dedicated a track to auxiliary impacts (e.g. 'lands', 'body hits', 'tumbles'). These tracks contained isolated granular gestures - usually using different combinations of pads (and corresponding sound model parameters) - that pertained to more complex events in the character's animation, such as collapsing onto the ground and shuffling limbs during transient states of locomotion. Four participants (P1,P2,P5,P6) consistently separated sequences of repetitive movement for locomotion from more transient states. Others used this track to only render particular key events. For example, P3 used this track to contrast the majority of limb impacts with the sound of the character's head hitting the wall during its descent at the end of scene 4, and its subsequent impact with the floor at the beginning of the next scene.

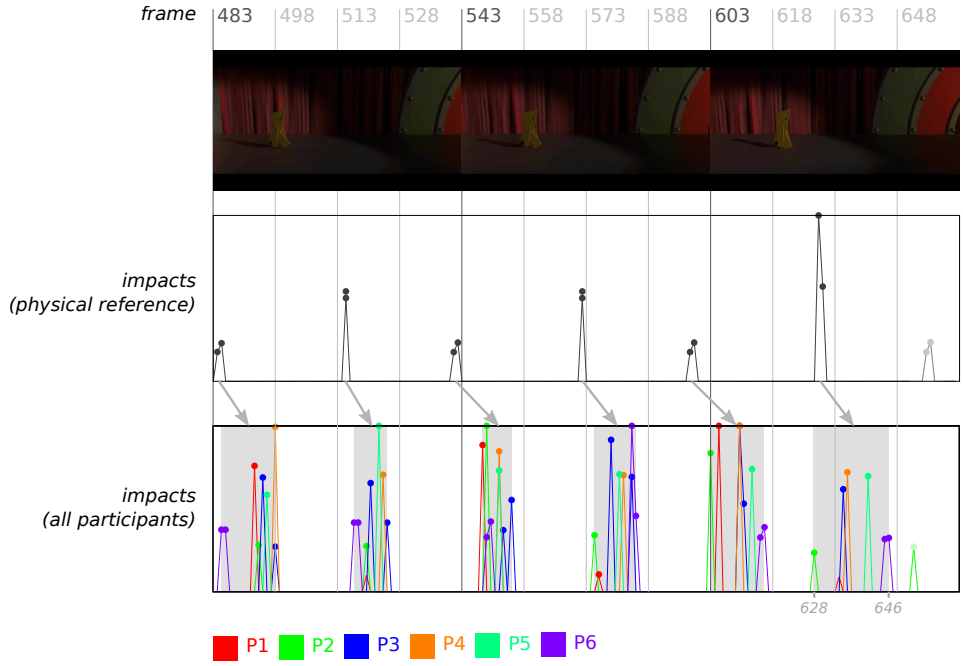


Figure 6.4: Performed impacts for footstep tracks across all participants during sequence of regular locomotion in Scene 2.

### 6.4.2 Omission and Exaggeration of Synchronisation to Physical Movement

Squeaks and scrapes were commonly used to mark key events in the figure's movement. For example, P1 used squeaks to mark transitions from standing to walking and vice versa (see Figure 6.2). At some points these gestures mirror the angular velocity extracted from the animation in both temporal placement and relative intensity. This is particularly evident at onsets and offsets of repetitive movement for locomotion, where it would seem that all physical movement aside from the first and last instance was *omitted* as a means of emphasising or punctuating the transitions. At other points they seem to be more distinct from the physical reference data, This is particularly evident in the third participant's soundtrack, where 'squeaks' were explicitly distinguished from 'vocals' across two separate tracks. As suggested by the label, the latter served the function of marking or projecting an emotional state of the character (as described in the screenplay) at a particular point of the narrative, for example 'getting a fright' when approaching a ledge in Scene 2.

While most participants associated a 'vocal' or 'anthropomorphic' quality with the squeaks, the majority was less explicit in the delineation between movement and

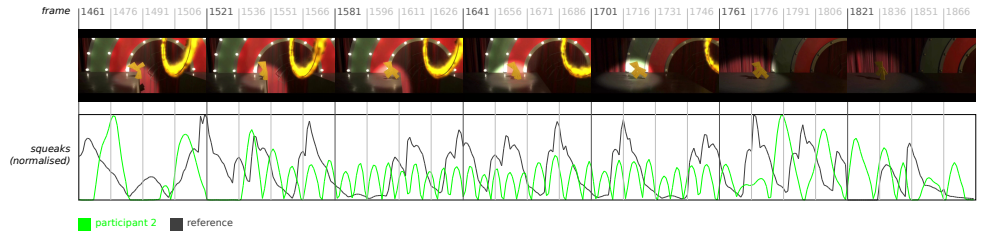


Figure 6.5: Exaggerated squeaks performed by P2 during sequence of character running away from ledge in Scene 2.

vocalisation. This resulted in tracks where key events were embedded within longer sequences of movement, manifesting themselves as *exaggerations* of physical movements or gestures that were inconsistent with the animation. An example is shown in Figure 6.5, where P2 performed rapid squeaks to accompany the sequence of the character running away from the ledge in Scene 2. This stands in clear contrast to both the physical reference data and performed sounds for other sequences of locomotion, which were more consistent with the animation. In most of these cases it is unclear whether these events take the specific role of a vocalisation, but nonetheless it would seem that these moments in the performed soundtracks serve one of two purposes:

- Marking contrasting or transitioning states of movement
- Expressing a meaningful point in the narrative that is not deducible from the objective movement alone (such as the character's internal state)

Such functional components of the soundtrack are commonly discussed by film sound theorists, where Chion (1994) would refer to the former as 'punctuating movement' and the latter as 'rendering' or 'added value'. The former commonly serves the function of guiding the viewer through the on-screen action, breaking it down into more psychologically digestible chunks. The latter can be applied to heighten emotional aspects of the image or bring to the fore elements that would otherwise not be present. These are intentional human elements of the soundtrack that are impossible to achieve in a direct mapping of physical movement to sound. Evidence for this can be gleaned in testimonies made by some participants upon viewing the physics-based reference soundtrack amongst human performed ones (see Section 6.2.6). While every participant rated it as being superior in terms of synchronisation and precision, two participants described the physics-based rendering as being 'too full' (P3) and another as 'sterile' (P1), even before being told that it was not performed by a human.

### 6.4.3 Deviations in Source-Sound Relationships

In the design of the experimental environment for this study, the sound models were chosen with particular source-sound relationships in mind. In other words, each impact model corresponded to one of the character's limbs and the sounds that it produced corresponded to its collisions against a surface. While the sounds were introduced to the participants as 'impacts', 'scrapes' and 'squeaks', no relationship to the animation was enforced, mentioned or verbally suggested - they were free to apply them in any way they wanted while performing the soundtrack.

As a result, sound models were not always identified with the source that they were originally designed to represent. The most commonly occurring example of this is the scrape model which was used to render rotational limb movements ('sliding') almost as frequently as scrapes along the surface (particularly when performed independently of the impacts).

More extreme deviations included the use of the scrape model to sonify air 'whooshing' past the character while jumping in the air (see Figure 6.6), and granular sequences of impacts as an almost musical accompaniment to the character's descent into the pit at the end of the fourth scene. Therefore sound models were not only interpreted as corresponding to different sources, but were repurposed by participants to correspond to a variety of sources.

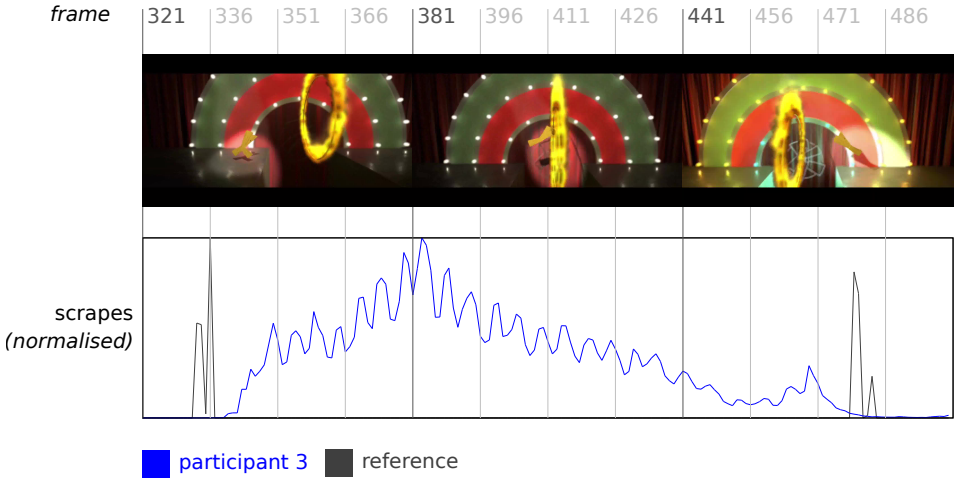


Figure 6.6: 'Whooshing' track performed with the scrape model by rubbing a finger back and forth on one of the pads for Scene 3 by P3.

Unlike in the direct physical mapping of movement to sound, the participants were able to choose which combination of sound sources to use to render physical movement, as well as how this 'mapping' could deviate over the duration of a sequence. Maybe

the most illustrative example of this can be found in the use of contrasting resonator settings to render complex limb impacts (e.g. collapsing upon landing). A more nuanced example is the omission of repeated sounds during locomotive movement in order to emphasise specific events (discussed above).

These stylistic choices are maybe more pronounced here due to the simplicity of the sound models and the resulting ambiguity of their source - for example, multiple participants associated scraping sounds with 'breath' (e.g. blowing into a glass bottle). The musical association of the sound models by some of the participants also signifies the perception of an abstract or non-existent sound-source relationship, potentially leaving more room for interpretation by the participant. This suggests a creative use of constraints on behalf of the participants, in that they have made the most out of what was initially considered to be a restricted sound palette. Most pertinently, this suggests that organising, punctuating or *rendering* the moving image through sound seems to have taken a primary role for most participants, who were willing to go to lengths of repurposing the available sonic palette in order to compose a more meaningful soundtrack.

### 6.4.4 Events vs Continuous Movement

The categorisation of sounds and the order in which the corresponding tracks were recorded for each scene mirrors the conventional Foley workflow to some extent, where the soundtrack is typically divided into *footsteps*, *moves* (or *cloth passes*) and *spots* (and recorded in this order) (Ament, 2009). As most films that incorporate Foley into the soundtrack involve a lot of human characters and movement, footsteps are seen as a fundamental element of Foley. Moves are equally essential - though taking on a secondary role - referring to the movement of the corresponding character's clothing. While footsteps are typically carefully synchronised on a frame by frame basis by a Foley editor after being performed, it is the Foley artist's responsibility to add expressive qualities through their performance. *Moves* on the other hand require less precise editing, following the rhythm of the footfall while serving as a supporting background layer, adding further character and depth.

Half of the participants used scrapes to render the rotational movements of the wooden limbs instead of or in addition to wood-on-surface scrapes (which was their sole purpose in the mapping of physical reference data). This can be explained by a desire to add something that resembles a move track, as opposed to the alternative function of the sound model which would bear a closer resemblance to 'scuffs', an element that would naturally be included in the performance of footsteps. Other participants instead used a *Squeaks* track to closely follow the footsteps.

Finally, *spots* refer to specific sounds that are either marked by a supervising sound editor or necessitated by the presence of a particular object or action on the screen

(e.g. lifting a cup). While such cues were intentionally omitted from the screenplay provided in this study, every participant dedicated at least one track to these kinds of actions; the most commonly occurring ones being 'body impacts' when the character lands on the floor.

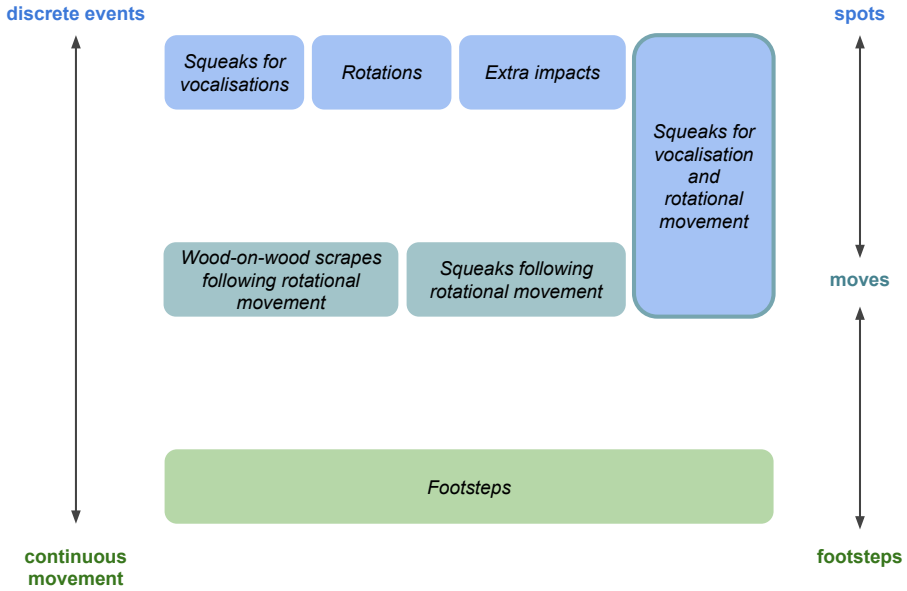


Figure 6.7: Organisation of sound categories according to their correspondence to Foley conventions (*footsteps,moves,spots*).

In Figure 6.7 a selection of sound categories identified by participants has been organised according to their correspondence to each of the three labeling conventions of *footsteps*, *moves* and *spots*. While performed tracks labeled as 'footsteps' are characterised by a fairly strict correspondence to the physical movement in the animation (collisions between limbs and a surface), tracks such as those labeled as 'rotations' or 'extra impacts' were constructed out of more isolated gestures corresponding to key events identified in the animation. Within this context of temporal granularity, the former could be said to correspond more closely to continuous movement imposed by the animation system, while the latter would be more easily structured as discrete events. Similarly to the footsteps, tracks corresponding to 'moves' followed the physical movement fairly strictly, but were more likely to omit or exaggerate particular sequences. As such, they might fall somewhere between the sonification of continuous movement and discrete events where the latter are demarcated *subtractively* (omission) or *additively* (exaggeration).

Understanding performed soundtracks in the context of continuous movement and

discrete events is helpful in obtaining a better understanding of how some of the stylistic choices made by participants could be integrated into an interactive system. As discussed in previous chapters, current integration strategies for computational models and recorded assets are very different in that the former assume a one-to-one mapping between physical movement and sound (like the physical reference track used in this study) while the latter is based on specific events and states extracted from the game engine. The soundtracks performed by participants contain a lot of complexity despite being based on a fairly constrained set of sound models, and therefore require a correspondingly complex integration strategy that transcends the possibilities afforded by solely event-based or physics-based approaches. This will be illustrated through a selection of short case studies of how particular aspects of the soundtrack relate to the physical reference data, and potential strategies for integrating them into an interactive system.

## 6.5 Case Studies

This section presents five brief case studies from the performed soundtracks with the aim of exploring potential integration strategies in an interactive context.

### 6.5.1 Footsteps during regular footfall

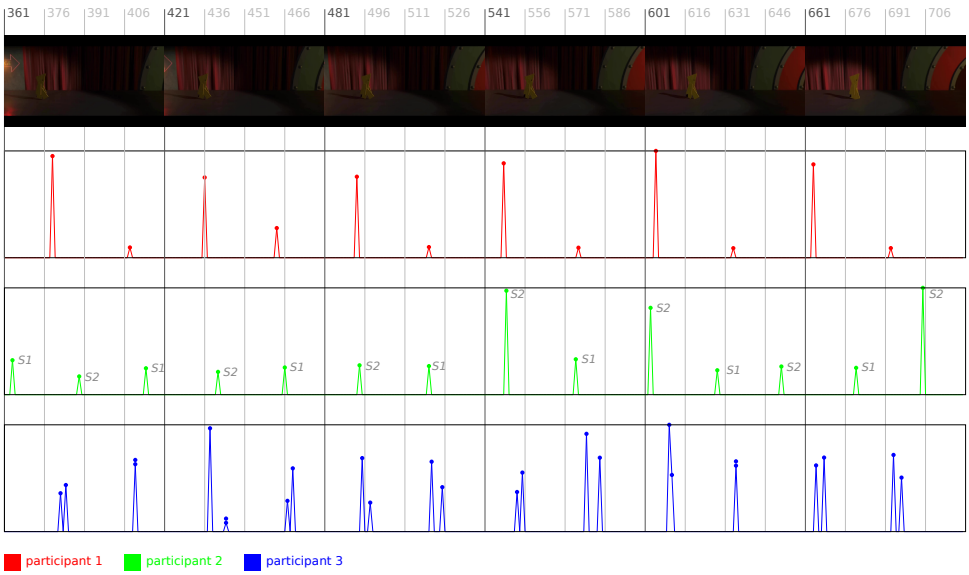


Figure 6.8: Performed footstep tracks for steady walking state in Scene 2 by three participants.

#### Observation

Figure 6.8 shows three different footstep tracks performed by P1, P2 and P3 for a steady period of walking during the second scene of the animation. The first thing to highlight here is that while there are clear errors in synchronisation of up to 15 frames (the animation has a frame rate of 60Hz), each performed impact is intended to correspond to the character's legs colliding with the ground at each step of the walk. Two of the three participants decided to distinguish between two different legs during the walk. While P1 used a single pad to perform this walk but alternated in intensity for each leg, P2 used two separate sensors where the corresponding sound models were set to slightly different sizes, resulting in alternating pitches. P3 used a single sensor but didn't differentiate between alternating legs. Instead two impacts were performed for each footfall with spacings and intensities varying arbitrarily. As mentioned in the previous section, the physical reference data also consists of two impacts per footfall (due to slight asymmetricities caused by the physics-based animation system) but have spacings of no more than two frames for this particular walking sequence.

#### Integration

Recreating these observed effects in a technical integration would require more information than is provided by the physical movement extracted from the animation. In the first two cases, the walking sequence would need to be expressed as singular 'steps' (as opposed to generic impacts) with a corresponding phase (left leg forward or right leg forward). There might be two strategies for formalising the third participant's performance. On one hand each 'step' extracted from the animation could correspond to a sequence of two impacts, where spacing and intensity vary arbitrarily according to a separate parameter. On the other hand, as the physics-based animation already consists of two impacts, some pseudo-random jitter could be applied to the incoming data adjusting the timing and intensity arbitrarily.

### 6.5.2 Body Impacts

#### Observation

Moments of the character collapsing onto the floor (e.g. upon termination of jumps) were commonly performed on a separate track. Figure 6.9 illustrates three different occurrences of the animated character terminating a jump in the first scene, as performed by P5. These segments stand in stark contrast to the participant's footstep track, which was performed using two sensors and a high level of precision. Impacts for these events were performed on up to five pads simultaneously (corresponding to different size and resonance settings) and bear little relation to the physical data.

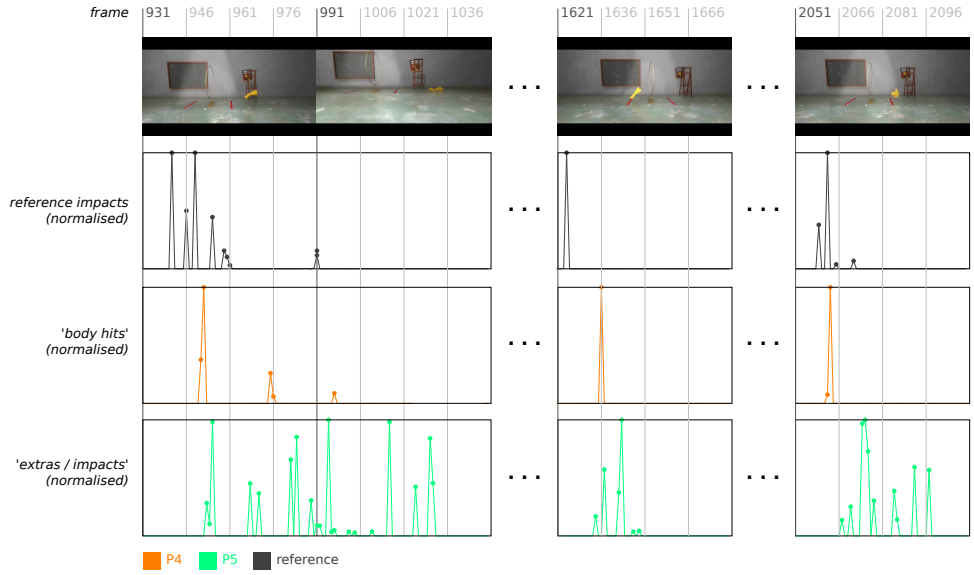


Figure 6.9: Impact data for 'Body Hits' performed by P4 and 'Extras' performed by P5 as the animated figure terminates jumps in Scene 1.

Each event contains more performed impacts than collisions present in the corresponding physics data. The duration, intensity and granularity of these gestures seem to correspond to the overall intensity of the corresponding physical movement, but the micro-structure of the events appears to be arbitrary, suggesting a singular gesture rather than a carefully synchronised sequence of impacts.

Other participants rendered these events with sequences of impacts that followed the physical movement more closely. An example of this can be found in the 'body hits' track performed by P4 (see Figure 6.9). These cases still contrasted with the corresponding footstep tracks through the use of different fixed parameters for the impact models.

### Integration

In all of the above cases, each 'body impact' can be regarded as a discrete event that is distinguishable from the more continuous and tightly-coupled synchronisation of the footsteps. In the first case, there is a very loose correspondence between physical movement and the resulting sound. The performed sound is perhaps better understood as a discrete gesture, where the underlying structure varies in duration, intensity and granularity depending on the perceived intensity of a given falling motion. A parameter corresponding to this could be extracted from the animation system by considering both physics data and animation states.

In cases where congruence to physical movement is less important the extracted parameter could be used to trigger or blend across various performances (as per the event-sound paradigm). When the animation system detects a 'body impact', the corresponding intensity of the fall is calculated based on available physics data and used to trigger corresponding performances.

In other cases the sequences would require following the physical movement more closely, while maintaining a certain level of irregularity. One approach might be to detect events on a more granular level of detail. For example, the overarching event of collapsing onto the ground could be structured into finer-grained sets of impacts (based on their physical intensity or temporal proximity.

### 6.5.3 Transitional States

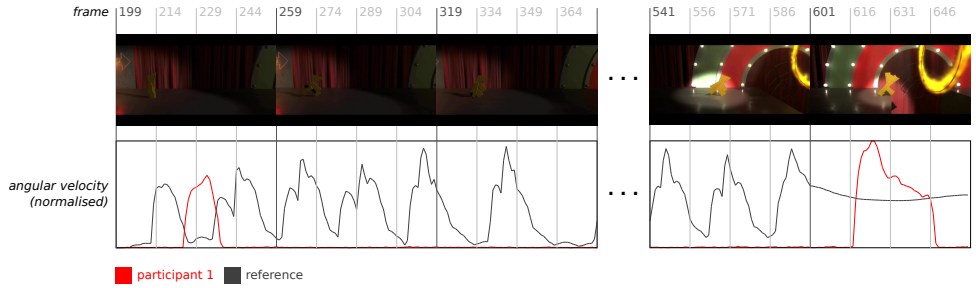


Figure 6.10: Punctuation of transitional state of character in Participant 1's 'squeak' track for Scene 4.

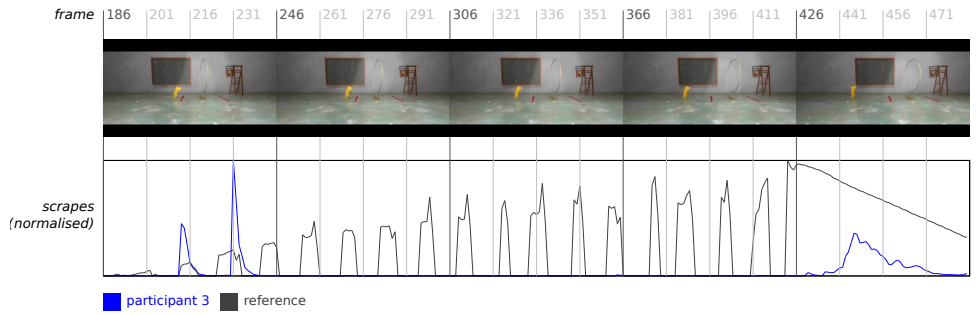


Figure 6.11: Punctuation of transitional state of character in Participant 3's 'slides' track for Scene 1.

#### Observation

As discussed above, a common feature of tracks following the rotational movement of the figure was the omission of repeated events in order to highlight transitional states of the animation. Figure 6.10 and 6.11 illustrate two participants' use of the squeak and scrape models to highlight transitions between idle and locomotive states. In each case the start and end of the locomotion is marked by a scrape or squeak gesture. In these two cases the performed gestures seem to follow the physical movement relatively closely. In a similar case discussed above (see Figure 6.2) the terminating squeak is more distinct and isolated.

#### Integration

The first two scenarios point to an integration strategy wherein repeated segments of continuous movements are simply omitted before being passed into the sound model during steady periods of locomotion.In other cases, an event-driven approach might be more appropriate, where transitional states are extracted from the animation system and trigger a gesture (which could vary depending on the corresponding walking style).

### 6.5.4 Whooshing

#### Observation

The soundtrack developed by P3 featured a unique use of the scraping model to emulate the sound of air rushing past the character while jumping across ledges and falling down the pit. This was performed using rapid back-and-forth movements on the interface, much like a *tremolo* gesture in violin performance. While the speed of the gestures remained constant throughout each sequence, the intensity was modulated based on the character's position. In sequences of jumping through the ring, the intensity of the gesture followed the general parabolic trajectory of the jump. During the falling sequence, the gestures increased in intensity as the sequence progressed, reaching maximum intensity at the end of the scene. An additional track was used (performed using a higher pitched setting of the scrape model) in order to exaggerate this effect, with the lower pitched sound only entering in the final two seconds of the scene.

The back-and-forth movement was applied as an attempt to overcome a constraint of the interface (the inability to perform uninterrupted scrapes due to the short length of the sensor). The participant stated that they would have ideally performed this sound with a continuously increasing and decreasing shape, but stated later that they were pleased with the result, despite it being unusual.

#### Integration

Automating these sequences would require extracting a value from the environment pertaining to the trajectory of the character's jumping motion. This could be based on the character's height above the ground during jumping sequences and the distance traveled in the air during falling sequences. For the participant's original intention of continuous sound this value could directly control the variable parameter of the scraping model, bypassing the constraints of the physical interface. During the falling sequence further mediation would be required in order to offset the introduction of the lower-pitched model instance, so that it is only introduced after a certain distance has been traveled. However, it is also possible to conceive of an integration strategy that imitates the tremolo gesture performed by the participant. A parametric model of this gesture could be developed based on oscillatory motion, where the corresponding amplitude is modulated based on the aforementioned value extracted from the game engine.

### 6.5.5 Dependencies on Narrative and Animation Class

#### Observation

Participants sometimes varied their synchronisation strategy depending on the narrative context or the character's style of walking. As noted above, P1 repeatedly used the squeaking model to mark the beginnings and endings of walking sequences. In Scene 5, however, this participant used a squeaking gesture on every footfall in order to emphasise the distinct walking style, or emotional state, of the character.

A similar case can be observed in running sequences performed by P2. While squeaks were normally performed for every footfall during running sequences in Scenes 1 and 4, a distinctly erratic sequence of squeaking sounds was performed during the last half of the second scene as the character runs away from the ledge. This corresponds to the 'frantic' and 'frightened' state of the character specified in the screenplay.

#### Integration

Both of these cases would require the game engine to switch between integration strategies depending on the animation class (e.g. walking or shuffling) or the narrative context (character is frightened or confident). This implies a hierarchical structure in the soundtrack. While a combination of physics and animation data would be required to recreate the individual walking styles (as discussed above), further information regarding the narrative context or locomotion style would be used to switch between each walking style (see Figure 6.12).

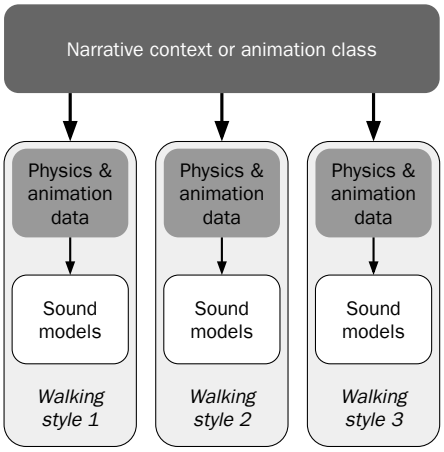


Figure 6.12: Hierarchical structure in context-dependent integration of contrasting walking styles.

### 6.5.6 Summary: Prospective Mediation Strategies

Table 6.9 presents an overview of the most common events and movements identified by participants, organised by their ability to be extracted from the animation system developed for this study. As illustrated in the previous overview of case studies, a higher-level description of the movement would be required in order to formalise sequences performed by participants. While this points to an event-driven approach to integration, the majority of sequences would still require a tight correspondence to the continuous physical movement represented by collisions, scrapes and angular velocity extracted from the underlying physics engine. This suggests that an integration strategy that takes these stylistic choices into account would require a hierarchical structure, wherein mappings between physical data and sound model parameters are nested within a higher level description of events. While physical data pertaining to continuous movement plays an important role, a layer of mediation is required in order to incorporate any of the stylistic decisions observed in the performed soundtracks.

## 6.6 Conclusion

### 6.6.1 Sources, Actions and Layers

Perhaps the most crucial finding made here is in the vertical organisation of the participant's soundtracks. In the physics-based integration, the soundtrack was structured according to physical events and corresponding sound models: surface collisions (impacts), surface slides (scrapes) and limb rotations (squeaks). While participants

| | |
|---|---|
| **Extractable from physics data** | limb collisions<br>floor scrapes<br>angular velocity of limbs |
| **Extractable from animation system** | footfall<br>jumping<br>landing<br>rotating (changing direction)<br>walking style<br>collapsing (ragdoll transition)<br>standing up (ragdoll transition)<br>mid-air state and velocity |
| **Not presently extractable** | narrative context (e.g. hitting wall after tripping over ledge)<br>emotional state of character (e.g. getting a fright)<br>somersaults<br>compound or accidental movements at state transitions (e.g. shuffling) |

Table 6.9: Events identified by participants in their soundtracks organised by their ability to be extracted from the interactive system.

usually performed one model at a time (maintaining separation across sound classes), the corresponding performances were categorised according to particular actions (e.g. 'footsteps', 'tumbles') with multiple layers of the soundtrack often relying on the same model. The synchronisation task required participants to pay close attention to the temporal unfolding of physical events, such as collisions, but the way in which these events were paired to sound depended on higher-level categorisations of movements and narrative states.

Parallels to Foley art conventions can be drawn in the participants' soundtracks, for example in the delineation between *footsteps* and isolated events, such as body impacts, resembling *spot effects*. With regard to their integration into an interactive audiovisual system, the conventional layers of *footsteps*, *moves* and *spots* can be placed along a spectrum ranging from continuous movements to events. For example, footsteps always required a precise temporal synchronisation to physical collisions, even if their performance entailed a clear stylisation of the physical events. In contrast, sounds corresponding to 'body impacts' and 'tumbles' were often based on isolated events identified by a given participant. In terms of temporal and physical accuracy, scrapes and squeaks seemed to play a secondary role. Performances with these models were more likely to diverge from the physical data in order to embellish or punctuate the character's movement.

### 6.6.2 Mediation

As noted above, physical data would be crucial in automating the majority of the effects observed in the participants' soundtracks (in the case of this animation). In the case of footsteps, each performed counterpart would ideally be fitted to coincide more or less perfectly with the collision data extracted from the animation. Similarly, squeaks and scrapes often corresponded closely to the combined angular rotation of the character.

However, it is also clear from the above observations that the physical data in itself would not suffice to formalise the majority of the audio-visual relationships performed by the participants. For example, distinguishing between 'body impacts' and 'footsteps' in the soundtrack would require obtaining higher level information from the animation system in order to deduce the character's state. Even in the precise synchronisation of footsteps, further information would often be required in order to distinguish between left and right legs and transient states of locomotion. Thus some *mediation* between the physical data and the variable parameters of the sound models is required in order to formalise these synchronisation strategies.

On one hand, such mediations are informed by higher-level information extracted from the virtual environment. For example, different approaches to rendering footsteps were taken depending on both the narrative context (e.g. the emotional state of the character) and the animation class (e.g. whether the character is walking or shuffling). On the other hand, relationships between data extracted from the animation and the performed soundtrack are often best understood by considering the physical gesture that was used. One of the most obvious examples of this was in the use of back-and-forth gestures on the touch-capacitive pads to render 'whooshing' sounds following the character's trajectory while jumping. While footstep tracks would have ideally been temporally fitted to the character's footfall, the underlying structure was often based on recurring gestural patterns. Similar co-dependencies between physical data and gestural patterns were observed in the performance of 'body impacts'.

This suggests that an ideal integration strategy would involve deducing a model of the performed gestures. In some cases it would be conceivable to produce an explicit parametric model based on the observed gestures (for example oscillatory movement to mimic the tremolo-like gestures for 'whooshing' sounds). Depending on the complexity of the desired soundtrack it may not be convenient or productive to design explicit parametric reductions of performed gestures. A promising solution may be found in the application of machine learning approaches, where complex temporal dependencies between gestural and animation data are learned based on examples provided by the performer. An alternative, albeit less organic approach might repurpose tools from graphical animation such as *state machines* and *blend trees* in order to dynamically switch and interpolate between synchronisation styles.

This will be discussed in more detail in the following chapter.

### 6.6.3 Constraints and Creative Opportunities

Finally, a more unexpected finding made in this study is in the way that sound models were repurposed by participants to emulate different sources. This ranged from the use of scrape models to simulate 'whooshing' sounds, to more subtle effects such as the use of the squeaking model to perform vocalisations. While many participants considered the range of the sound models to be limited and different in comparison to their typical approach with physical props, each of them found the range of sounds sufficient to perform an adequate soundtrack for the animation. Participants mitigated perceived limitations of the sound model by imposing new source-sound relationships and employing contrasting playing styles for each action category.

It is also worth noting here the musicality associated with the perceived simplicity of the sound models and the pad interfaces. Participants frequently made comparisons to 'old-school' animation soundtracks where every action would be recreated through music. Perhaps this perceived limitation in the sound palette led to a more exaggerated approach to the synchronisation task than they would normally employ on a regular Foley stage. It would be interesting to explore whether similar strategies are employed by Foley artists when presented with more realistic sound models and interaction strategies.

# Chapter 7

# Discussion

One of the first observations made at the start of this research project was that current implementations of CGA inherently assume a direct mapping between action and sound. In other words, sound models are typically designed to produce audible output in response to a clearly defined front-end of time-varying parameters that in turn accord with streams of values from the corresponding application (e.g. a virtual environment or a sonically augmented object). These can be based on physical or perceptual principles – for example, a wind model could be driven by a 'wind intensity' parameter or by an 'air velocity' parameter, depending on the amount of physical detail that is in the underlying simulation. In either case, this imposes a requirement of genericity on the sound model that makes it difficult to tailor the output to specific situations in a target application. As discussed in Chapter 3, the *behaviour* of the model is typically ingrained within the sound model, either implicitly as part of a numerical simulation (in the case of a source model) or explicitly, in the form of a *behavioural abstraction* (in the case of a signal model). This is technically advantageous as it enables a process-driven and object-oriented approach to sound design, as envisioned by Farnell (2008). For example, models can be instantiated multiple times with parameter settings that are unique to the corresponding objects they are attached to. Compound models can be generated on the fly by combining multiple lower-level models according to the computational constraints of a given context.

However, when viewed in the light of established sound design practice and aesthetics (as discussed in the fields of film sound theory and electro-acoustic music) a number of shortcomings to this approach become apparent. The most pertinent issue addressed in this thesis is that sound design relies on contradiction or *asynchronicity* between sound and the action or movement that it corresponds to in order to elicit meaning (Pudovkin, 1985) (Chion, 1994). If the soundtrack simply mimicks what is

already observable in another sensory channel then this limits its potential for expressivity (in other words, its ability to disclose *added value*). The resulting soundtrack is not necessarily devoid of meaning: as Chion (1994) points out, absolute redundancy between the sound and image is not possible because there is no such thing as an objectively ideal soundtrack. However, the ability to creatively intervene in the way that sound and image interact (compliment and contradict each other) is lost in the case where a sound model is integrated by means of a one-to-one mapping. In other words, subjective intentionality in the creation of the soundtrack is reduced.

The aim of this work has been to explore ways of reintroducing this intentionality by means of performative real-time interaction with sound models. This has been approached in two aspects of the computational soundtrack. First, in the performative design of behaviours by controlling lower-level (or *timbre-led*) perceptual descriptions of the sound model (as opposed to a physics-based parameterisation). Second, by exploring the performed synchronisation of behaviour to complex visual movement. The following findings have been made as a result of these endeavours:

- Instead of encoding behaviour within the sound model, it can be performed separately, even when controllable parameters do not correspond closely to physical processes.
- Suitable sensor-mapping-model couplings make it possible to perform custom sound sequences in real-time that are as believable as ones that were manually designed according to reference material, while maintaining adequate levels of controllability, nuance and range.
- On one hand, this results in more expressive output from the computational model that is tailored to a given narrative context; on the other hand, it complicates the process of *integrating* the model into an interactive application.
- Even in the case of a physics-based front-end, when given the chance to perform, experts demonstrated divergences between represented movement and sound model parameters that would be difficult or impossible to automate using existing integration paradigms.

The remainder of this chapter serves as a discussion of these findings with a particular emphasis on the implications of integrating observed behaviours into an interactive context. Section 7.1 explores some of the technical considerations involved in integrating the most basic form of performed sound model behaviour as implemented in Chapters 3 and 4. Parallels are drawn to graphical animation, where motion of virtual object representations is typically expressed as separate interpolatable sequences of data. Section 7.2 reviews the key findings of the final synchronisation study, highlighting ways in which basic interpolation techniques fall short, rendering both the

event-sound paradigm and the one-to-one mapping of movement to sound inadequate for the meaningful integration of performed soundtracks. Finally, Section 7.3 serves as a reflection on performing CGA and opens the discussion to broader issues including unexpected findings that warrant further investigation in future work. Some of the fundamental assumptions underlying this research (and the broader fields surrounding CGA) are challenged, most pertinently, the notion that a sound model should correspond to a virtually represented source.

## 7.1 Looking Ahead: Technical Considerations

One of the first considerations in the introduction of performance into the design and control of computational sound models has been in the level of abstraction at which real-time control is introduced. Central to this process was the separation of performed behaviour from a fixed signal chain constituting a computational sound model. As discussed in Chapter 3 performed behaviour can correspond to physical parameters (in the case of physics-based models) or perceptual ones in the case of timbre-led models.

On one hand real-time performance of sound models using physical sensors enables the rapid design of behaviours that would be hard or difficult to achieve numerically, while simultaneously leveraging the performer's ability to encode expressive nuances. On the other hand, expressing the model's behaviour as data (as opposed to procedurally encoding it within a behavioural model)also means that the typical approach of mapping dynamic properties of an interactive environment directly to model parameters is no longer valid when it comes to its integration. As illustrated in Figure 7.1, the current approach to integration is to map real-time parameters from the interactive application to top-level parameters of the sound model. For example, in the case of a physics-based door creaking model for a game an *angular velocity* parameter would be extracted from the game engine and mapped directly to a corresponding parameter in the sound model. The sound model reacts to parameter changes and produces audible output that changes according to an implicit behavioural model.

The above solution is convenient as the virtual door produces responsive sound following a very straight-forward integration process.[1] However, the human element of performance is lost when relying on a purely computational behavioural layer.

Chapter 4 presented a *performable model* of a creaking door, and an evaluation study demonstrated that participants could easily perform complex context-dependent behaviours without the need to modify the sound model's internal structure. The study did not account for means of integrating these performances back into

[1] Some basic transformations might be applied in order to match parameter ranges, but the mapping is still *one-to-one*.

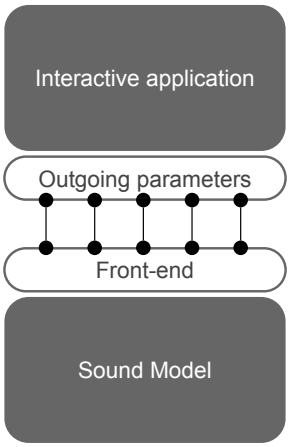


Figure 7.1: Conventional integration of a computational sound model.

an interactive application, but a few approaches can be considered. As illustrated in Figure 7.2 an intermediary layer would be required in order to manage the playback of performed sequences. In the case of the radio play that was developed as part of the study, performances were simply played back at the corresponding cues, just as one would trigger ordinary waveforms. Applying this same procedure to games brings the integration process back to the event-sound paradigm that is prevalent in current game engines and middleware platforms: the game triggers events (e.g. 'door opening') which in turn trigger the playback of a recorded sequence. While there could be some advantages to such a rudimentary implementation (e.g. reduced memory footprint for a large amount of sequences), by itself it carries all the same disadvantages as the conventional sample-based approach: interactivity is limited and a large number of sequences is required in order to obtain a diverse range of output.

While simple playback of sequences is clearly limited, a more powerful approach can be found in the dynamic processing of performance data. Because the recorded sequences are defined at a much lower rate than audio sampling rate and correspond to *states* of the sound model rather than waveforms, they accommodate much more flexibility in the way they are processed. Unlike audio waveform data, sequences can be played back at faster or slower rates without producing any artifacts in the output. Similarly, blending two sequences results in a true intermediary stage as opposed to the audio-equivalent of cross-fading, which causes two waveforms to be superimposed on one another.

Parallels can be drawn to graphical animation. Here, pre-existing geometrical models of objects or characters are augmented with transformation points along which delegated sections of the geometry can be rotated, transformed or scaled. Each trans-

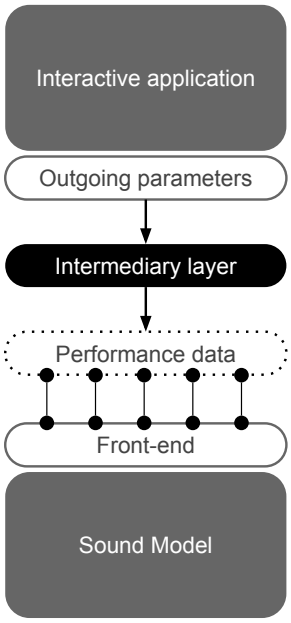


Figure 7.2: Integration of performed data for a computational sound model.

formation point is typically given a set of constraints, narrowing the total *degrees of freedom (DOFs)* for that particular point. By setting corresponding values for each DOF, a pose can be created. An animation is created by interpolating between a set of poses that are distributed across arbitrary points in a temporal sequence. This process is commonly referred to as *key-frame animation* (where each pose corresponds to a key-frame) (Multon et al., 1999).

More elaborate operations can be performed on sets of key-frames, comprising the broader field of *parametric animation*. For example, in a technique referred to as *motion blending*, complex transitions can be achieved by interpolating between two or more animations. Parameter values are weighted according to a variable blending coefficient and summed as the sequences unfold over time. This can be employed to make a character transition from a walking to a running state, for example. A more complex parametric animation might consist of hierarchical sets of poses and sequences, where a higher-level parameter space allows the animator to combine and animate multiple states (e.g. transitioning from walking to running while waving a hand).

There are many parallels to be drawn to CGA, particularly to the case where performed behavioural elements are separated from a fixed model. While parameter sequences are not defined using a set of static 'poses' the same processes of

interpolation or 'blending' can be applied to generate compound behaviours out of a relatively small set of base sequences. A set of higher level parameters can be defined to interpolate across different weightings and to control the playback of the resultant compound behaviour. Some of these ideas were implemented in the *FoleyDesigner* project - a hardware and software prototype developed over a five-month period based on concepts presented in this thesis (see Appendix D). In one of the demonstrations of the developed prototype, two meta-parameters are defined to control a creaking door model similar to the one presented in Chapter 4. A *screechiness* parameter blends between two different performances while an *angle* parameter acts as a playhead for both animations (normalised in time). See Appendix D.7 for more details and Appendix E.4 for a supplementary video.

While these techniques present a promising avenue for future work, they also imply a workflow that can become very complex and eventually starts veering away from the benefits that have been associated with human performance in CGA. For example, many of the stylistic choices observed in Chapter 6 occurred serendipitously as a result of experimentation in a performative context. It may be difficult to preserve some of the resultant qualities in the process of developing an explicit formalisation.

A more performance-centric approach may be found in applying machine learning techniques to the integration process. Mapping-by-demonstration methods proposed by Fiebrink et al. (2009) and Françoise et al. (2014) can be applied to learn complex mappings between sound parameters and dynamic states of an interactive environment by performing sounds to a set of example sequences. A powerful concept outlined by Fiebrink and Caramiaux (2017) is that input to learning algorithms can be modified in real-time, as part of the creative process. This idea is proposed in the context of musical performance, but could easily transfer to the development of a computational soundtrack. In such a scenario, the sound designer takes on the role of a Foley artist, performing sound effects to visual sequences as they unfold in real-time (for example, while another person interacts with the virtual environment). While the resultant mapping layer is obfuscated or *hidden* (as opposed to the previous example where mappings are explicitly defined by means of parametric hierarchies), the real-time implementation would rely on similar principles of interpolation and temporal warping.

Future work in this area is required to get a clearer understanding of the limitations of interpolating across data-representations of parameter sequences. For example, interpolation of parameter states in FM synthesis can lead to unnatural or unexpected trajectories due to inherent complexities in the signal process (Yee-King and Roth, 2008). Transition points and phasing considerations (in the case of oscillatory motions) will also need to be addressed (this has received a lot of attention in the field of graphical animation Multon et al. (1999)). A further question regards the use of

complex control layers in the performance and integration of sequences. Should the sensor output be recorded and interpolated while preserving corresponding control layers in the real-time context, or can a similar effect be obtained by interpolating the processed parameter inputs for the sound model?

## 7.2 Towards True Asynchronicity

Parametric interpolation techniques introduced above have the potential to offer new workflows for CGA that are based on high-level manipulation of performed sequences. On one hand they focus the design task on manipulating custom-performed data in an on-the-fly manner, encouraging a rapid prototyping and design process centred around performance. On the other hand, designers can develop highly complex behaviours that would be hard or impossible to express programmatically within a self-contained model (i.e. what was referred to as *behavioural abstraction* in Chapter 3). Despite all this, however, this strategy does not inherently consider many of the expressive devices in the audio-visual image that have been explored in the previous two chapters of this thesis.

When experts were given the chance to perform the soundtrack to complex sequences of movement with fairly loose constraints (aside from the requirement of using a pre-determined set of computational models), several intentional divergences from a direct integration of the models could be identified. These divergences, or *asynchronicities* have been categorised into two classes: *horizontal* asynchronicity, where movement-sound relationships diverged over time, often in relation to the given narrative context or animation state, and *vertical* asynchronicity, where source-sound relationships delegated by the participants often differed from the presumed function of each sound model.

### 7.2.1 Horizontal Structure

Temporal divergences are perhaps best understood by considering points at which actions performed through the sound models consistently *converge* with events or movements in the animation, what Michel Chion might refer to as *synch points* Chion (1994). Findings from the study have shown that these corresponded not only to physical data extracted from the animation system, but also to higher level states of the animation and properties relating specifically to the narrative context of the motion. Furthermore, there was some hierarchy evident in the way various categories of sound corresponded to different classes of synchronisation points.

Performed tracks that were associated with 'footsteps' corresponded closely to physics data, and in fact were often referred to as being the most difficult to synchronise performatively. In fact, it would seem that these performances would have

benefited greatly from having been automatically aligned to physical events (collisions between the character's legs and the floor). The interesting patterns observed here were not directly in the way that individual collisions performed by participants diverged in time from those extracted from the simulated movement. Instead they were found in the way that sounds were grouped *within* the occurrence of a single synch point and *across* the evolution of multiple synch points. An example of the former case would be instances where each individual event classed as a 'footstep' consistently comprises of multiple impacts occurring in a formalisable pattern. An example of the latter would be a broader overarching trajectory (such as a narrative arc) affecting the way that individual events (and their corresponding micro-structures) are modulated over time.

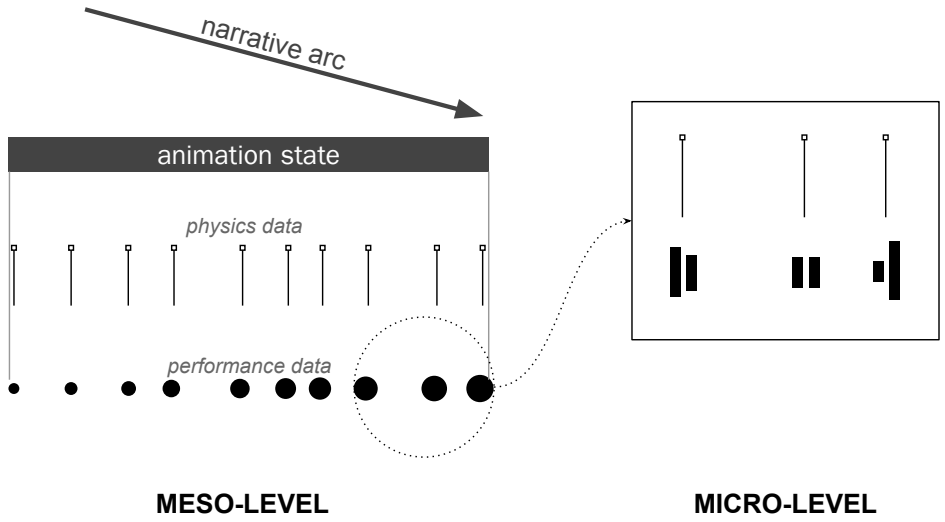


Figure 7.3: Meso and Micro-level dependencies of performed soundtracks.

A strong parallel can be drawn to Godøy's concept of *gestural-sonorous objects* (Godøy, 2006). Here, Schaeffer's definition of the *sound object*, is recontextualised to the domain of instrumental (musical) gesture. Godøy's stance is that musical sequences undergo a process of *recoding* on a gestural level upon listening (and, by extension, upon performing), whereby musical fragments are delineated through a process of *chunking* (Miller, 1956). Thus sequences of sound can be inspected on a *meso-level* and a *micro-level*, where the latter corresponds to the structure of individual fragments and the former to higher-level sequences of fragments. Figure 7.3 illustrates how this concept could be applied to the kinds of audio-visual scenarios described above. While a detailed discussion of the perceptual processes underlying this theory falls outside the scope of this discussion, this marks a potentially important

parallel between expressive audio-visual relationships and the study of musical gesture. Therefore a fruitful avenue for future research would be a more detailed study of how physical gesture corresponds to performed sound synchronisation (for example by analysing motion capture data of Foley artists). Organising performed soundtracks based on physical gesture might be a useful way of achieving a balance between the event-sound paradigm of current game audio integrations and temporally-dependent structures observed in the synchronisation study.

Mapping strategies based on supervised machine learning (*mapping by demonstration*) developed by Francoise et al. (2012) take a lot of inspiration from Godøy's frameworks, where hierarchical and nested structures are integrated into the learning model. It would be interesting to investigate how well these and related technologies perform at learning complex structures like the ones observed here. In doing so it is important to keep in mind the types of real-time information that need to be extracted from the interactive application. These extend beyond purely physical data and include information corresponding to the animation states (e.g. whether a character is walking or jumping) as well as the overarching narrative structure. In the latter case, one should keep in mind that what constitutes a narrative is often ambiguous in an interactive scenario and may therefore be hard to deduce unless it is explicitly defined in the application.

### 7.2.2 Vertical Structure

Another interesting outcome of the synchronisation study was in the way that participants used sound models to represent various elements of the animated sequences.

In all cases individual takes were recorded according to categories of movement (e.g. footsteps, floor impacts, etc.) rather than sound classes (i.e. collisions, scrapes, squeaks). On one hand this reflects the practice of Foley artists, where the soundtrack is by convention commonly split into footsteps, moves (or 'cloth passes') and spots. But it also reinforces the notion of object-wise structuring as discussed above. For example, while 'footsteps' and 'floor impacts' are two distinct categories of actions, they both correspond to the same source-sound pairing (limbs colliding against floor). However, the ways in which performed gestures correspond to movement is markedly different - this is not only observable in the designation of distinct categories but also in the structure of the performed data (see Section 6.5.2). This is a clear example of how a *one-to-one* mapping of physical data to a sound model precludes the intentional demarcation of *actions*.

The prioritisation of action over source emulation is commonly addressed in SID, where sound design for interactive commodities is commonly approached with an explicit focus on gestural affordances of objects (e.g. (Altavilla et al., 2013; Delle Monache et al., 2008b)) and libraries of sound models have been developed and organised cor-

respondingly (Delle Monache et al., 2010).[2]

A more obvious divergence from the physics-based reference soundtrack was in the use of particular sound models to render entirely different sources from those that they were intended for. Examples included use of the scraping model to render the sound of wind rushing past the character while mid-air, and the impact models to emulate the sound of limbs rattling against each other (despite no limb-on-limb collisions being present in the simulation). On one hand this can be attributed to the constrained set of resources (or 'virtual props'). The topic of constraints, and their creative exploitation, will be addressed shortly in the following section. But in the meantime there are some other interesting observations to be made here regarding the vertical structure of the performed soundtrack. The participant that chose to render the character's limbs rattling against one another only applied this effect during the sequence of the character falling down a hole, in a climatic moment in the narrative. Similarly, the effect of rushing wind was only applied to scenes where the character was jumping across a stage (many more jumping sequences occur in the 'rehearsal room' scene). In each case these effects fulfilled a particular function in the narrative that the corresponding participants felt a need to emphasise using a contrasting source-sound relationship. While it is highly likely that these participants would have used a different object to generate these sounds had they been on a real Foley stage, this indicates moments where the *function* of the sound overrides its ability to provide a detailed emulation of the rendered source.

Further to this, these sequences exemplify the dynamic omission or addition of sources in order to emphasise particular aspects of the animation or narrative. Some participants commented that the physical reference track was 'too full' or 'too busy'. One interpretation of these claims is that the situation where all sources are constantly producing sound (as would be the case in a physically realistic rendition) precludes the opportunity to punctuate the moving image through the dynamic orchestration of sound sources.

A more subtle subversion of source-sound relationships can be seen in the use of squeaks to provide an element of vocalisation for the character. Four out of the six participants commented on the 'vocal character' of the squeaks, but only one of them assigned two separate tracks for 'squeaking' and 'vocalising'. Moments at which the gestures on the squeak model corresponded to vocalisation instead of rotational movements could sometimes be deduced by observing the data, but the delineation between the two was typically unclear from looking at the data alone. The use of the voice to add rich meaning to the description of a sound through imitation is a common

[2]However, these applications are very different from the scenarios investigated here, in that they are typically constrained to a very particular set of interactions with a clearly specified function. In this sense there isn't the same issue of achieving a balance between rendering sources and expressive asynchronicity, and a close coupling between action and sound is in most cases required in order for the interactive sound to fulfill its function in the context of the given commodity.

phenomenon observed in everyday communication between humans and particularly in children during playful interactions (Dumaurier et al., 1982). The ability to achieve similar effects in the comparatively basic interaction with the crank interface (i.e. one degree of freedom) is an interesting proposition that warrants further study. Recent research being undertaken in vocal sketching (see for example (Delle Monache et al., 2015; Piccolo and Rocchesso, 2016) and the European Union funded *SkAT-VG* project (Rocchesso et al., 2015)) sheds increasing light on some of the psychological devices underlying these phenomena and their application in sound design.

## 7.3 Reflections on Performing CGA

The interfaces employed in the synchronisation study followed the enactive approach (Essl and O'Modhrain, 2006) (in order to more easily accommodate a direct comparison to data extracted from the animation). The touch-capacitive and force-sensitive pads could be struck with a finger tip in order to produce collision sounds, or rubbed to perform scrapes. The crank interface resembled a wooden corkscrew that produced squeaks corresponding to the angular velocity at which it was turned. It is noteworthy how every participant showed a clear preference for the latter. Participants commented on the clarity of the intended interaction and one participant noted that it was much more like a 'Foley tool' than like a 'musical tool' (referring to the pads in the latter case). Its strong resemblance to an object that is likely to produce this sound (including the wooden handle and concealment of any electronic components) was probably a strong factor here. In contrast, the pads were perceived as being more unwieldy, not producing the desired response to physical force and likened more to a musical instrument (e.g. a keyboard or a xylophone) than something that would typically be used in a Foley studio. This raises the question of what role *realism* in the physical interface plays on its perceived suitability for performing sounds. Parallels can be drawn to the notion of presence in virtual reality - referring to the point at which the user suspends their sense of disbelief and perceives virtual objects or environments as being real. Future work in the only recently emerging field of Virtual Reality Musical Instruments (Serafin et al., 2016) could be highly relevant here.

Limitations were not just perceived in the interface but also in the sounds that they produced. For example, some participants associated the scraping sounds with a 'breathy' quality, resembling the sound of blowing into a hollow cavity rather than scraping an object across a surface. While this is partly due to the simplicity of the sound models, it would be worth investigating whether the perceived realism of the interface has an impact on the way sources are associated with the sound it produces.

Despite the comparatively negative response to the pads, some of the most innovative uses of the corresponding impact and scrape models were performed on this

interface. For example, one participant used a tremolo-like back and forth motion to render the sound of wind rushing past the character (also described above). Other participants performed clusters of sounds to accompany sequences of the character collapsing onto the floor by tapping all of the pads in an arbitrary order.

A key observation to make from the above is that participants found creative solutions when confronted with the limitations of the interfaces.[3] These discoveries were often made in the process of completing the synchronisation task rather than in the initial phase of discovery. The inclusion of wind rushing sounds and rattling limbs occurred as a result of impromptu experimentation with the interface and was not pre-determined upon first viewing of the animation or initial familiarisation with the interface. In other words, ambiguities (in the context of realism) in the sound models and the corresponding interfaces resulted in more experimentation and on-the-fly decision making in the process of producing a soundtrack.

As suggested by Gaver et al. (2003), the ambiguity of the interface can be an important resource here. In Foley, conventional (acoustic) tools used to perform sound often bear little resemblance to the objects represented on screen, and are often discovered through creative exploration and serendipity (see Section 5.2.3). Similarly, the most faithful physical sound model of the visual object might not be preferable to a less accurate but more versatile model.

In this case the resulting sound quality would not have been ideal for a final product, and the observed process could be likened to *sonic sketching* (Delle Monache et al., 2015), in other words, exploring potential options through a limited but familiar apparatus before producing a finalised soundtrack. However, tying this line of thought back to the previous discussion of integration techniques raises the question: does impromptu performance and decision-making need to be confined to an intermediary stage of exploration, or can it play a more central role in the design, performance and integration of CGA?

It is of course left up to future research and development to see what technologies will underlie a successful integration system for performance-driven CGA. It could resemble tools that are currently being used in the graphical animation industry or it could be based on supervised machine learning techniques - or on other ground-breaking technologies that have not been considered in this research or are yet to be discovered. Regardless of the underlying technology, there is a arguable case to be made for on-the-fly workflows that incorporate performance deeply into the design and integration process.

In an analogy to hyperrealistic paintings Puronas (2014) likened the current sample-based approach prevalent in sound design to sonic taxidermy and suggests instead ap-

[3] A parallel can be drawn here to studies conducted by Gurevich et al. (2010) and Zappi and McPherson (2014), where intentional constraints imposed on a musical interface resulted in higher degrees of creative exploration and appropriation.

proaching the soundtrack as a 'sonic painting' through computational techniques. In keeping with this metaphor, performance-led systems (given the appropriate integration solution) have the potential to equip the sound designer with a sonic *paintbrush*, allowing the workflow of designing sound to exist in a perpetual state of creativity and experimentation.

As computational resources increase it will become more feasible to incorporate increasingly realistic physical source models. Even here, on-the-fly integration techniques that are centred around performance would open the doors to soundtracks that extend beyond physical realism and incorporate all the innovation and creativity that can be found on the Foley stage.

However, the source-sound relationship need not be a realistic one, and little stands in the way of new aesthetic sensibilities arising from soundtracks that prioritise behaviour and process over the aggregation of emulated sources. In the history of sound design and music this can be witnessed in the orchestral scores of early animated cartoons, and in the emergence of radiophonic music and spectralism. In the visual arts, impressionism, stop-motion film and visual montage are just few examples of how creative appropriations of technology have resulted in novel and culturally foundational aesthetic forms.

## 7.4 Summary

Sound design is a creative process and an artistic endeavour that has always been shaped by the tools at hand. Examples range from the limited audio capturing equipment that gave rise to Foley art, to the creative manipulation of recorded sound that led to radiophonic music and now, the hand-full of audio middleware platforms that the majority of modern games rely on. It is important to consider future tools that give rise to new opportunities, even if (or especially when) the implications of their eventual usage is unknown. This applies particularly to CGA, still in its infancy, where any future creative opportunities will be dependent on the technical toolchains that facilitate their integration into interactive media.

# Chapter 8

# Conclusions and further work

This thesis has shed light on two key aspects in the stylistic enhancement of computationally generated audio through the incorporation of human performance.

A new class of *timbre-led models* has been proposed to facilitate the performative design of behavioural sequences as external components. By exposing a set of perceptual features as the central dynamic parameter space for a model, behavioural sequences can be manually composed or performed in real-time.

A range of stylistic and creative limitations have been exposed pertaining to the use of CGA to sonify visual movement on a purely physical basis. Based on these findings, future integration strategies for CGA would benefit from a combined consideration of physical variables, narrative context and higher-level events and states as used in current sample-based approaches.

Section 8.1 provides a summary of the contributions made in this thesis. Section 8.2 presents some reflections and directions for future work.

## 8.1 Summary of contributions

### 8.1.1 Separation of Behavioural Components from the Sound Model

An extension to practical synthesis methods by Farnell (2008) has been proposed, whereby the behavioural properties of a sound model are explicitly detached from an underlying signal chain, resulting in what has been referred to here as a *timbre-led model*. By exposing a subjectively meaningful perceptual parameter space, behaviours can be composed as discrete sequences externally to the model rather than integrating them as a fixed process. A design strategy for iteratively exposing a parameter space has been proposed and implemented in the development of a timbre-led creaking door

model. This procedure incorporates the generation of behavioural sequences to match reference sounds as a central part of the process, with the intention of developing a contrasting and linearly independent perceptual parameter space.

### 8.1.2 Application of Mapping Strategies to Perform Sound Effects in Timbre Space

Unlike behavioural models that are most naturally controlled using *enactive* interaction strategies, timbre-led models consist of a parameter space more closely resembling synthesisers implemented in musical instruments. A contrasting set of mapping strategies inspired by DMI research was applied to control the timbre-led creaking door model. Results of a user study indicated that each of these mapping strategies facilitated the performance of sound effects for narrative contexts that were deemed satisfactory by the performer. Interestingly, the most favoured control strategy was also found to be the most distinguishable from a listener's point of view. One interpretation of this is that what *feels right* to the performer does not necessarily *sound right* in a listening context.

Issues surrounding difficulty and virtuosity (commonly referred to in the context of a 'learning curve') were identified by many of the participants, particularly for the most complex of the three implemented control strategies. While a sustained learning process (leading to increasing levels of musical skill and virtuosity) is often a desirable feature of a musical instrument (Jordà, 2005), this is less likely to be the case in the performance of sound effects. This is reflected in results from the survey of Foley artists (presented in Section 5.2), where high levels of timbral range and *source appropriability* (see Sections 4.4.4 and 5.2.3) were deemed important factors in the selection of physical objects to be performed with.

### 8.1.3 Experimental Environment for Comparing Performed Soundtracks to Physical Reference Data

The objective of studying asynchronicity in CGA led to the development of a complex experimental environment that could be re-used and developed in further research. Several limitations were found following the synchronisation study described in 6. Reactions to the interfaces employed in the final study suggest that physical realism play an important factor in the performance of source models (or behavioural models). The crank, which closely resembled a physical mechanism associable with the corresponding sound was favoured by every participant in the study, while the force and touch sensitive blocks were deemed more unnatural and 'musical'. The physical animation system was crucial in providing a set of physical reference data and successfully portrayed emotionally varied movement that was suitable in engaging the

participants. Nonetheless, some elements of the experimental environment could be improved in future iterations. For example, the movement in the animation was considered to be unusually complex in relation to more common animation practices in television and cinema, particularly in sequences of locomotive movement.

### 8.1.4 Stylistic Strategies in the Synchronisation of CGA to Complex Movement

Observations from the synchronisation study described in Chapter 6 illustrate how intentional deviations in the sound-image relationship were indeed present and in most cases a direct consequence of real-time performance. While microscopic deviations in temporal synchronisation (i.e. frame accuracy of collision sounds) were generally not intended or desired (as preempted by survey responses presented in Section 5.2), interesting temporal patterns arose in conjunction with transitory states of movement and other events not explicitly defined in the physical data. A common strategy was to *omit* or *exaggerate* sounds pertaining to particular sequences of physical movement in order to punctuate the general movement or an overarching narrative structure. This points to a hierarchical temporal structure that takes a combination of continuous movement, discrete events or states, and narrative context into account.

Furthermore, participants often categorised performed sequences (takes) according to *actions* rather than *sources*. For example, 'impacts' were distinguished from 'footsteps' and 'squeaks' were distinguished from 'vocalisations'. This suggests that future integration strategies might consider a mediation layer corresponding to action classes when controlling sound models.

Another important observation was that source-sound mappings chosen by participants were not always the same as the sound models suggested, with sound models often serving the rendition of diverse sources. To some extent this is due to the limited range of sound models and inherent source ambiguity due to their simplicity (e.g. scraping sounds were often confused with 'breath' sounds). On the other hand this illustrates a phenomenon of creative appropriation, which could play an important role in future implementations of CGA. In similar vein to the exclusively orchestral soundtracks of early animated cartoons, it shows that accurate representations of sources are not always necessary as long as they support the rendition of meaningful behaviours.

## 8.2 Reflections and Further Work

The following subsections provide some reflections on key areas of interest with a focus on further avenues of research.

### 8.2.1 Behaviour and Timbre

The proposed design methodology for developing timbre-led models was only applied to a single case study of a creaking door. While observations from the user study presented in Section 4.5 suggest that there is potential in this approach, it would be worth evaluating the design procedure to a wider range of sound effects or sources. As discussed in Section 4.2, the sound of a creaking door was a convenient case to study due to the flexibilities afforded by its continuous nature and the diverse variety of timbres associated with it. It would be particularly interesting to examine how well the framework of performable timbre-led models extends to other sound effects or sources with more discontinuous or transient features. For example, how would one approach designing and performing a timbre-led model that produces door slamming sounds?

Future work in the development of such models would also have to address the problem of parameter dimensionality. The creaking door study was limited to three of most commonly occurring parameters, but strategies for approaching a larger dimensional parameter space are yet to be explored in this context. One approach might be to iteratively perform different groups of parameters, progressively adding timbral detail in a workflow that resembles visual sculpture. Approaches to reducing or collapsing the parameter space must be careful to avoid undesired artifacts, as observed in implementations by Tubb and Dixon (2014) and Yee-King and Roth (2008).

As noted by Pachet and Aucouturier (2004) any definition of a timbre space is a highly subjective and what constitutes a perceptually meaningful and linearly independent set of parameters is likely to vary across different people. This subjectivity is of course inherent in the proposed design methodology and thus some implementations are likely to only be useful to the person that has designed the model. It would be interesting to explore this workflow in a collaborative sound design context, involving multiple individuals in the iterative evaluation and exposure of perceptual dimensions. Here, many links could be drawn to *cooperative sound design* approaches studied recently by Erkut et al. (2016) in the context of vocal sketching.

### 8.2.2 Performance Interfaces and Mapping Strategies

It is interesting that the physically-inspired control layer received the lowest ratings for satisfaction and other metrics in the first user study. While this was attributed to difficulty (the requirement to overcome a long learning curve) the tangibility of the interface may also play a crucial role. A similar (albeit more simplified) mediation layer coupled with a more explicit physical affordance was the most favourable interface in the final study. It would be interesting to investigate whether the listening strategy applied has an effect on these ratings. In the case of the timbre-led model, perfor-

mances are guided by the control of perceptual parameters to compose the impression of physical behaviour. Precision may be a higher factor here, where performers are more likely to be considering micro-structures within a discrete behavioural sequence. In the case of performing Foley the focus shifts to larger hierarchies of actions and thus an enactive approach that mimics everyday interactions may be more suitable.

### 8.2.3 Integration Strategies and Workflow

Maybe the most logical continuation of this research lies in the development and evaluation of integration strategies in order to automate performed sequences. Chapter 7 outlined a number of potential solutions. Many parallels can be drawn to established techniques in the field of graphical animation, which rely on dynamically interpolating between pre-designed sequences of movement (or, in this case, pre-performed sequences of behaviour). This was explored to some extent in the *FoleyDesigner* project (see Appendix D), where meta-parameters were used to blend between performed sequences. Much more work can be done in this direction, for example in the use of state machines to develop a more complex range of behaviours based on a larger set of sequences. Integrating such a system directly into existing animation tools and collaborative design workflows with animators are further avenues worth exploring here.

Observations from the final study point to an integration strategy were sequences performed with physical gestures correspond to a hierarchical set of events and trajectories. Machine learning techniques that take temporal hierarchies into account (e.g. Francoise et al. (2012)) may be a fruitful avenue to explore. Such a *mapping-by-demonstration* approach would stand in contrast to the use of techniques borrowed from animation. On one hand, it leverages performance as the central activity in the integration process, eliminating the requirement to explicitly define a dynamic system based on discrete sequences. On the other hand, a hidden learning layer might limit the ability to make detailed adjustments to the learned synchronisation structure, or require an impractical amount of examples.

Creative appropriation of the sound models and interfaces observed in the synchronisation study are also worth noting here. It would be interesting to explore integration strategies that encourage experimentation with the provided tools to obtain more imaginative and unpredictable results. Much in line with propositions by Puronas (2014), it is not inconceivable that future processes of designing and integrating computational soundtracks could be analogous to painting, where a combination of calculated decisions, experimentation (e.g. 'happy accidents') and fast iterations contribute to a unique and expressive result.

## 8.3 Closing Remarks

In a sense observations made in this thesis may shed some light on the question of why CGA has not yet been adopted largely in the games industry. On one hand, sample-based methods offer a guaranteed level of quality that is easily achievable by established means. More pertinently, however, it is worth considering whether asynchronous effects achieved by the static nature of the samples outweigh the benefits of a direct mapping of continuous parameters for a comparatively sterile physical model.

As interactions in virtual environments become increasingly complex and unpredictable, CGA techniques are likely to play a more prominent role in games and virtual reality applications. An intermediary solution might be a calculated combination of conventional and computational techniques. Eventually, tools and workflows for CGA would benefit from a more deeply integrated approach that reacts to both discrete events and continuous movement in natural and organic ways. The work presented in this thesis has shown that human performance, while offering a useful lens through which to appreciate aesthetic shortcomings of CGA, may be an important means of achieving these goals.

# Bibliography


Andrew Allen and Nikunj Raghuvanshi. Aerophones In Flatland: Interactive Wave Simulation of Wind Instruments. *ACM Transactions on Graphics*, 34(4):134:1–134:11, July 2015.

Alessandro Altavilla, Baptiste Caramiaux, and Atau Tanaka. Towards Gestural Sonic Affordances. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2013.

V.T. Ament. *The Foley Grail: The Art of Performing Sound for Film, Games, and Animation*. Taylor & Francis, 2009.

Julian Anderson. A provisional history of spectral music. *Contemporary Music Review*, 19(2):7–22, January 2000.

Newton Armstrong. *An enactive approach to digital musical instrument design*. PhD thesis, Princeton University, 2006.

Maribeth Back and D. Des. Micro-narratives in sound design: Context, character, and caricature in waveform manipulation. In *Proceedings of the International Conference on Auditory Display*, 1996.

Stefano Baldan, Stefano Delle Monache, Davide Rocchesso, and Hélene Lachambre. Sketching Sonic Interactions by Imitation-Driven Sound Synthesis. In *Proceedings of the Sound and Music Computing conference*, Hamburg, Germany, 2016.

Natasha Barrett. Kernel Expansion : A Three-Dimensional Ambisonics Composition Addressing Connected Technical, Practical and Aesthetical Issues. *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics.*, 2010.

Ej Berger. Friction modeling for dynamic system simulation. *Applied Mechanics Reviews*, 55(6):535, 2002.

Frédéric Bevilacqua, Bruno Zamborlin, Anthony Sypniewski, Norbert Schnell, Fabrice Guédy, and Nicolas Rasamimanana. Continuous realtime gesture following and

recognition. In *Gesture in embodied communication and human-computer interaction*, pages 73–84. Springer, 2009.

Stefan Bilbao. *Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics*. Wiley Publishing, 2009a.

Stefan Bilbao, Brian Hamilton, Alberto Torin, Craig Webb, Paul Graham, Alan Gray, Kostas Kavoussanakis, and James Perry. Large Scale Physical Modeling Sound Synthesis. In *Proceedings of the Stockholm Musical Acoustics Conference/Sound and Music Computing Conference*, Stockholm, 2013.

Stefan Bilbao, Alberto Torin, Paul Graham, James Perry, and Gordon Delap. Modular physical modeling synthesis environments on GPU. In *Proceedings of the International Computer Music Conference*, 2014.

Stefan D Bilbao. *Numerical sound synthesis finite difference schemes and simulation in musical acoustics*. John Wiley & Sons, Hoboken, NJ, 2009b.

Peter Birkholz, B. J. Kröger, and P. Birkholz. A survey of self-oscillating lumped-element models of the vocal folds. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pages 47–58, 2011.

Niels Böttcher and Stefania Serafin. Design and evaluation of physically inspired models of sound effects in computer games. In *Proceedings of the 35th International AES Conference: Audio for Games*, 2009.

W. F. Brace and J. D. Byerlee. Stick-Slip as a Mechanism for Earthquakes. *Science*, 153(3739):990, August 1966.

Matthew Burtner. The Metasaxophone: concept, implementation, and mapping strategies for a new computer music instrument. *Organised Sound*, 7(02), January 2003.

Claude Cadoz. Instrumental gesture and musical composition. In *Proceedings of the International Computer Music Conference*, pages 1–12, Cologne, Germany, 1988.

Claude Cadoz. Supra-Instrumental Interactions and Gestures. *Journal of New Music Research*, 38(3):215–230, September 2009.

Claude Cadoz, Annie Luciani, and Jean L. Florens. CORDIS-ANIMA. A modeling and simulation system for sound and image synthesis. The general formalism. *Computer Music Journal*, 17(1):19–29, 1993.

Baptiste Caramiaux, Patrick Susini, Tommaso Bianco, Frédéric Bevilacqua, Olivier Houix, Norbert Schnell, and Nicolas Misdariis. Gestural embodiment of environmental sounds: an experimental study. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2011.

A. Chaigne and V. Doutaut. Numerical simulations of xylophones. I. Time-domain modeling of the vibrating bars. *J. Acoust. Soc. Am.*, 101(1):539–557, January 1997.

Luciano Chessa. *Luigi Russolo, Futurist: Noise, Visual Arts, and the Occult.* University of California Press, Oakland, CA, 2012.

M. Chion. *Guide des objets sonores: Pierre Schaeffer et la recherche musicale.* Bibliothèque de recherche musicale. Buchet-Chastel Editions, Paris, France, 1983.

Michel Chion. *Audio-Vision: Sound on Screen.* Columbia University Press, New York, NY, 1994.

Hanwook Chung, Hoon Heo, DooYong Sung, Yoonchang Han, and Kyogu Lee. Modeling and Real-Time Generation of Pen Stroke Sounds for Tactile Devices. In *Proceedings of the 49th AES International Conference: Audio for Games*, London, 2013.

Eric Clarke. *Ways of Listening: An Ecological Approach to the Perception of Musical Meaning.* Oxford University Press, Oxford, UK, 2005.

Karen Collins. *Game sound an introduction to the history, theory, and practice of video game music and sound design.* MIT Press, Cambridge, MA, 2008.

Karen Collins. *Playing With Sound: A Theory of Interacting with Sound and Music in Video Games.* MIT Press, Cambridge, MA, 2013.

Perry R. Cook. Physically Informed Sonic Modeling (PhISM): Synthesis of Percussive Sounds. *Computer Music Journal*, 21(3):38–49, 1997.

Carlos C. De Wit, H. Olsson, K.J. Astrom, and P. Lischinsky. Dynamic Friction Models and Control Design. In *American Control Conference*, pages 1920–1926, San Francisco, CA, 1993.

S. Delle Monache, D. Devallez, C. Drioli, F. Fontana, S. Papetti, P. Polotti, and D. Rocchesso. *Sound synthesis tools for sound design.* Deliverable, 2008a.

Stefano Delle Monache, Stefano Papetti, Pietro Polotti, and Davide Rocchesso. Sonically augmented found objects. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Genova, Italy, 2008b.

Stefano Delle Monache, Pietro Polotti, and Davide Rocchesso. A toolkit for explorations in sonic interaction design. In *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*, Piteå, Sweden, 2010.

Stefano Delle Monache, D. Rocchesso, S. Baldan, and D. A. Mauro. Growing the practice of vocal sketching. In *Proceedings of the 21th International Conference on Auditory Display*, Graz, Austria, 2015.

Christopher Dobrian and Daniel Koppelman. The 'E' in NIME: musical expression with new computer interfaces. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 277–282, 2006.

Elisabeth Dumaurier, Bernadette Céleste, and François Delalande. *L'enfant du sonore au musical*. Buchet/Chastel, Paris, France, 1982.

Pierre Dupont, Vincent Hayward, Brian Armstrong, and Friedhelm Altpeter. Single state elastoplastic friction models. *IEEE Transactions on Automatic Control*, 47 (5):787–792, 2002.

Sergei Eisenstein. *Nonindifferent Nature*. Cambridge University Press, Cambridge, UK, 1987.

Inger Ekman and Michal Rinott. Using Vocal Sketching for Designing Sonic Interactions. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. ACM, New York, NY, 2010.

Jean Epstein. Slow-Motion Sound. In Elisabeth Weis and John Belton, editors, *Film Sound: Theory and Practice*. Columbia University Press, New York, NY, 1985.

Cumhur Erkut, Davide Rocchesso, Stefano Delle Monache, and Stefania Serafin. A Case of Cooperative Sound Design. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM Press, 2016.

G. Essl and S. O'Modhrain. Scrubber: an interface for friction-induced sounds. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 70–75, 2005.

Georg Essl and Sile O'Modhrain. An enactive approach to the design of new tangible musical instruments. *Organised Sound*, 11(03):285, November 2006.

Georg Essl, Stefania Serafin, Perry R. Cook, and Julius O. Smith. Theory of banded waveguides. *Computer Music Journal*, 28(1):37–50, 2004.

Andy Farnell. *Designing Sound*. MIT Press, Cambridge, MA, 2008.

Andy Farnell. Behaviour, Structure and Causality in Procedural Audio. In Mark Grimshaw, editor, *Game Sound Technology and Player Interaction: Concepts and Development*. IGI Global, 2011.

Andy Farnell. Sonarchy in the UK: is sound design a rebellious teenager? *The New Soundtrack*, 4(2):89–102, September 2014a.

Andy Farnell. Procedural Audio Theory and Practice. In *The Oxford Handbook of Interactive Audio*. Oxford University Press, 2014b.

Rebecca Fiebrink and Baptiste Caramiaux. The Machine Learning Algorithm as Creative Musical Tool. In *Oxford Handbook on Algorithmic Music*. Oxford University Press, Oxford, UK, 2017.

Rebecca Fiebrink, Dan Trueman, and Perry R. Cook. A metainstrument for interactive, on-the-fly machine learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, volume 2, page 3, 2009.

Jules Francoise, Baptiste Caramiaux, and Frédéric Bevilacqua. A Hierarchical Approach for the Design of Gesture-to-Sound Mappings. In *Proceedings of the 9th Sound and Music Computing Conference (SMC)*, 2012.

Jules Françoise, Norbert Schnell, Riccardo Borghesi, and Frédéric Bevilacqua. Probabilistic Models for Designing Motion and Sound Relationships. In *Proceedings of the 2014 International Conference on New Interfaces for Musical Expression*, London, UK, 2014.

William W. Gaver. What in the world do we hear? an ecological approach to auditory event perception. *Ecological Psychology*, 5:1–29, 1993.

William W Gaver, Jacob Beaver, and Steve Benford. Ambiguity as a resource for design. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 233–240. ACM, 2003.

Rolf Inge Godøy. Gestural-Sonorous Objects: embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound*, 11(02):149, 2006.

Rolf Inge Godøy. Images of Sonic Objects. *Organised Sound*, 15(01):54, March 2010.

Rolf Inge Godøy, A. R. Jensenius, and Kristian Nymoen. Chunking by coarticulation in music-related gestures. In *8th International Gesture Workshop, Bielefeld*, pages 25–27, 2009.

Kristian Gohlke, David Black, and Jörn Loviscach. Leveraging behavioral models of sounding objects forgesture-controlled sound design. In *Proceedings of the fifth*

*international conference on Tangible, embedded, and embodied interaction*, pages 245–248, 2011.

Michael Gurevich, Paul Stapleton, and Adnan Marquez-Borbon. Style and constraint in electronic musical instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2010.

Christian Heinrichs and Andrew McPherson. Mapping and Interaction Strategies for Performing Environmental Sound. In *IEEE VR Workshop: Sonic Interaction in Virtual Environments (SIVE)*, Minneapolis, MN, 2014.

Christian Heinrichs and Andrew McPherson. Performance-Led Design of Computationally Generated Audio for Interactive Applications. In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 697–700. ACM Press, 2016.

Christian Heinrichs, Andrew McPherson, and Andy Farnell. Human performance of computational sound models for immersive environments. *The New Soundtrack*, 4 (2):139–155, September 2014.

Thomas Hermann and Helge Ritter. Listen to your data: Model-based sonification for data analysis. *Advances in intelligent computing and multimedia systems*, 8: 189–194, 1999.

Lejaren Hiller and Pierre Ruiz. Synthesizing Musical Sounds by Solving the Wave Equation for Vibrating Objects: Part 1. *J. Audio Eng. Soc*, 19(6):462–470, 1971.

Olivier Houix, Stefano Delle Monache, Hélène Lachambre, Frédéric Bevilacqua, Davide Rocchesso, and Guillaume Lemaitre. Innovative Tools for Sound Sketching Combining Vocalizations and Gestures. In *Proceedings of the Audio Mostly 2016*, pages 12–19, 2016.

Daniel Hug. Towards a hermeneutics and typology of sound for interactive commodities. In *Proceedings of the CHI Workshop on Sonic Interaction Design*, pages 11–16, 2008.

Daniel Hug. Investigating Narrative and Performative Sound Design Strategies for Interactive Commodities. In *Proceedings of the 6th International Conference on Auditory Display*, CMMR/ICAD'09, pages 12–40, Berlin, Heidelberg, 2010.

Andy Hunt and Marcelo M. Wanderley. Mapping performer parameters to synthesis engines. *Organised Sound*, 7(02), January 2003.

Andy Hunt, Marcelo M. Wanderley, and Matthew Paradis. The importance of parameter mapping in electronic instrument design. *Journal of New Music Research*, 32(4):429–440, 2003.

Lea Jacobs. *Film Rhythm after Sound: Technology, Music, and Performance*. University of California Press, Oakland, CA, 2015.

David A. Jaffe and Julius O. Smith. Extensions of the Karplus-Strong Plucked-String Algorithm. *Computer Music Journal*, 7(2):56, 1983.

Alexander Refsum Jensenius. To Gesture or Not? An Analysis of Terminology in NIME Proceedings 2001-2013. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2014.

Sergi Jordà. Instruments and Players: Some Thoughts on Digital Lutherie. *Journal of New Music Research*, 33(3), 2005.

Matti Karjalainen, Vesa Välimäki, and Tero Tolonen. Plucked-String Models: From the Karplus-Strong Algorithm to Digital Waveguides and beyond. *Computer Music Journal*, 22(3):17–32, 1998.

Dean Karnopp. Computer Simulation of Stick-Slip Friction in Mechanical Dynamic Systems. *Journal of Dynamic Systems, Measurement, and Control*, 107(1):100–103, March 1985.

Kevin Karplus and Alex Strong. Digital Synthesis of Plucked-String and Drum Timbres. *Computer Music Journal*, 7(2):43, 1983.

Chris Kiefer, Nick Collins, and Geraldine Fitzpatrick. HCI methodology for evaluating musical controllers: A case study. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2008.

Michael Kubovy. Should we resist the seductiveness of the space:time::vision:audition analogy? *Journal of Experimental Psychology: Human Perception and Performance*, 14(2):318–320, 1988.

Michael Kubovy and Michael Schutz. Audio-Visual Objects. *Review of Philosophy and Psychology*, 1(1):41–61, 2010.

Michael A. Lee and David Wessel. Connectionist models for real-time control of synthesis and compositional algorithms. In *Poceedings of the International Computer Music Conference*, pages 277–280, San Jose, CA, 1992.

Guillaume Lemaitre and Davide Rocchesso. On the effectiveness of vocal imitations and verbal descriptions of sounds. *The Journal of the Acoustical Society of America*, 135(2):862–873, February 2014.

Marc Leman. Musical gestures and embodied cognition. In *Journées d'informatique musicale (JIM-2012)*, pages 5–7, Mons, Belgium, 2012.

Peter Lennox. *The Philosophy of Perception in Artificial Auditory Environments: Spatial Sound and Music.* PhD thesis, University of York, 2004.

Peter Lennox and Anthony Myatt. Perceptual cartoonification in multi-spatial sound systems. In *Proceedings of the International Conference on Auditory Display*, Budapest, Hungary, 2011.

James Leonard, Claude Cadoz, Nicolas Castagné, Jean-Loup Florens, and Annie Luciani. A Virtual Reality Platform for Musical Creation: GENESIS-RT. In *Proceedings of the International Symposium on Computer Music Modeling and Retrieval*, pages 346–371. Springer, 2013.

V. LoBrutto. *Sound-on-film: Interviews with Creators of Film Sound.* Sound-on-film: Interviews with Creators of Film Sound. Praeger, 1994.

Claudy Malherbe, Joshua Fineberg, and Berry Hayward. Seeing light as color; hearing sound as timbre. *Contemporary Music Review*, 19(3):15–27, January 2000.

Max V Mathews and Joan E Miller. *The technology of computer music.* M.I.T. Press, Cambridge, Mass., 1974.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264 (5588):746–748, December 1976.

M.E. McIntyre and J. Woodhouse. Friction and the bowed string. *Wear*, 113(1): 175–182, December 1986.

Andrew McPherson. TouchKeys: Capacitive Multi-Touch Sensing on a Physical Keyboard. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Ann Arbor, MI, 2012.

Andrew McPherson and Victor Zappi. An Environment for Submillisecond-Latency Audio and Sensor Processing on BeagleBone Black. In *Audio Engineering Society 138th Convention*, Warsaw, Poland, 2015.

Lucas Mengual, David Moffat, and Joshua D. Reiss. Modal Synthesis of Weapon Sounds. In *Audio Engineering Society Conference: 61st International Conference: Audio for Games.* Audio Engineering Society, 2016.

Dylan Menzies. Virtual Intimacy: Phya as an Instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Genova, Italy, 2008.

George A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 63(2):81–97, March 1956.

Ali Momeni and Cyrille Henry. Dynamic independent mapping layers for concurrent control of audio and video synthesis. *Computer Music Journal*, 30(1):49–66, 2006.

Giulio Moro, Astrid Bin, Robert H. Jack, Christian Heinrichs, Andrew P. McPherson, and others. Making high-performance embedded instruments with Bela and Pure Data. In *Proceedings of the International Conference on Live Interfaces*, Sussex, UK, 2016.

Martin James Morrell, Christopher A. Harte, and Joshua D. Reiss. Queen Mary's "Media and Arts Technology Studios" Audio System Design. In *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

William Moss, Hengchin Yeh, Jeong-mo Hong, Ming C. Lin, and Dinesh Manocha. Sounding liquids: Automatic sound synthesis from fluid simulation. *ACM Transactions on Graphics*, 2010.

Eoin Mullan. Physical Modelling for Sound Synthesis. In Mark Grimshaw, editor, *Game Sound Technology and Player Interaction: Concepts and Developments*. IGI Global, September 2010.

Christian Müller-Tomfelde and Tobias Münch. Modeling and sonifying pen strokes on surfaces. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland, December*, 2001.

Franck Multon, Laure France, Marie-Paule Cani-Gascuel, and Gilles Debunne. Computer animation of human walking: a survey. *The journal of visualization and computer animation*, 10(1):39–54, 1999.

Tristan Murail. After-thoughts. *Contemporary Music Review*, 19(3):5–9, January 2000.

Damian T. Murphy, Chris JC Newton, and David M. Howard. Digital waveguide mesh modelling of room acoustics: Surround-sound, boundaries and plugin implementation. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx)*, 2001.

James F. O'Brien, Perry R. Cook, and Georg Essl. Synthesizing Sounds from Physically Based Motion. In *Proceedings of ACM SIGGRAPH 2001*, pages 529–536. ACM Press, August 2001.

Sile O'Modhrain. A framework for the evaluation of digital musical instruments. *Computer Music Journal*, 35(1):28–42, 2011.

Francois Pachet and Jean-Julien Aucouturier. Improving timbre similarity: How high is the sky. *Journal of negative results in speech and audio sciences*, 1(1):1–13, 2004.

Alan Del Piccolo and Davide Rocchesso. Non-speech Voice for Sonic Interaction: a catalogue. *Journal on Multimodal User Interfaces*, 11(1):39–55, July 2016.

Michael David Plumpe, Thomas F. Quatieri, and Douglas A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *Speech and Audio Processing, IEEE Transactions on*, 7(5):569–586, 1999.

Daniel Pressnitzer and Stephen McAdams. Acoustics, psychoacoustics and spectral music. *Contemporary Music Review*, 19(2):33–59, January 2000.

Vsevolod Pudovkin. Asynchronism as a Principle of Sound Film. In *Film Sound: Theory and Practice*. Columbia University Press, New York, NY, 1985.

Vytis Puronas. Sonic hyperrealism: illusions of a non-existent aural reality. *The New Soundtrack*, 4(2):181–194, September 2014.

Zhimin Ren, Hengchin Yeh, and Ming C. Lin. Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)*, 32(1):1, 2013.

Curtis Roads. *The Computer Music Tutorial*. MIT Press, Cambridge, MA, USA, 1996.

Davide Rocchesso and Frederico Fontana, editors. *The sounding object*. Mondo estremo, Firenze, 2003.

Davide Rocchesso, Guillaume Lemaitre, Patrick Susini, Sten Ternström, and Patrick Boussard. Sketching sound with voice and gesture. *interactions*, 22(1):38–41, 2015.

Joseph Butch Rovan, Marcelo M. Wanderley, Shlomo Dubnov, and Philippe Depalle. Instrumental gestural mapping strategies as expressivity determinants in computer music performance. In *Proceedings of Kansei - The Technology of Emotion Workshop*, Genova, Italy, 1997.

Luigi Russolo. The art of noises. In Barclay Brown, editor, *The art of noises*. Pendragon Press, New York, NY, 1913.

P. Schaeffer, C. North, and J. Dack. *In Search of a Concrete Music*. California studies in 20th-century music. University of California Press, Oakland, CA, 2012.

S. Schedin, C. Lambourg, and A. Chaigne. Transient Sound Fields from Impacted Plates: Compariston Between Numerical Simulations and Experiments. *Journal of Sound and Vibration*, 221(3):471 – 490, 1999.

John C. Schelleng. The Bowed String and the Player. *The Journal of the Acoustical Society of America*, 53:26, 1973.

Rod Selfridge, Joshua D. Reiss, Eldad J. Avital, and Xiaolong Tang. Physically Derived Synthesis Model of an Aeolian Tone. In *Audio Engineering Society Convention 141*, September 2016.

Stefania Serafin. *The Sound of Friction: real-time models, playability and musical applications*. PhD thesis, Stanford University, 2004.

Stefania Serafin, Matthew Burtner, Charles Nichols, and Sile O'Modhrain. Expressive controllers for bowed string physical models. In *Proceedings of the International Conference on Digital Audio Effects*, Limerick, Ireland, 2001.

Stefania Serafin, Federico Avanzini, and Davide Rocchesso. Bowed string simulation using an elasto-plastic friction model. In *Proceedings of the Stockholm Music Acoustics Conference*, pages 95–98, Stockholm, Sweden, 2003.

Stefania Serafin, Amalia de Götzen, Niels Böttcher, and Steven Gelineck. Synthesis and control of everyday sounds reconstructing Russolo's Intonarumori. In *Proceedings of the Conference on New Interfaces for Musical Expression*, Paris, France, 2006.

Stefania Serafin, Cumhur Erkut, Juraj Kojs, Niels C. Nilsson, and Rolf Nordahl. Virtual Reality Musical Instruments: State of the Art, Design Principles, and Future Directions. *Computer Music Journal*, 40(3), 2016.

Stephen Sinclair, Marcelo M. Wanderley, Vincent Hayward, and Gary Scavone. Noise-free haptic interaction with a bowed-string acoustic model. In *Proceedings of the World Haptics Conference (WHC)*, pages 463–468, Istanbul, Turkey, 2011.

Denis Smalley. Spectromorphology: explaining sound-shapes. *Organised sound*, 2(2): 107–126, 1997.

Tamara Smyth and Julius O. Smith III. Creating sustained tones with the cicada's rapid sequential buckling mechanism. In *Proceedings of the Conference on New Interfaces for Musical Expression*, pages 1–4, Singapore, 2002.

D. Sonnenschein. *Sound Design*. Michael Wiese Productions, 2001.

David Sonnenschein. Sound Spheres: A Model of Psychoacoustic Space in Cinema. *The New Soundtrack*, 1(1):13–27, March 2011.

D. Stowell, A. Robertson, N. Bryan-Kinns, and M.D. Plumbley. Evaluation of live human–computer music-making: Quantitative and qualitative approaches. *International Journal of Human-Computer Studies*, 67(11):960–975, November 2009.

Dan Stowell. *Designing Sound in SuperCollider*. Wikibooks, The Free Textbook Project, 2012.

Tapio Takala and James Hahn. Sound rendering. In *ACM SIGGRAPH Computer Graphics*, volume 26, pages 211–220, 1992.

A. Tarkovsky and K. Hunter-Blair. *Sculpting in Time: Reflections on the Cinema*. University of Texas Press. University of Texas Press, 1987.

Randy Thom. Designing a Movie for Sound, 1999. Full text-version at: http://www.filmsound.org/articles/designing_for_sound. htm. [Retrieved on 1 March 2017].

Etienne Thoret, Mitsuko Aramaki, Christophe Bourdin, Lionel Bringoux, Richard Kronland-Martinet, and Solvi Ystad. Synchronizing Gestures with Friction Sounds: work in progress. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research*, Marseille, France, 2013.

Robert Tubb and Simon Dixon. A Zoomable Mapping of a Musical Parameter Space using Hilbert Curves. *Computer music journal*, 38(3), 2014.

Kai Tuuri. Gestural attributions as semantics in user interface sound design. In *Proceedings of the international conference on Gesture in Embodied Communication and Human-Computer Interaction*, Bielefeld, Germany, 2009.

Jean-Frederic Vachon. Avoiding Tedium - Fighting Repetition in Game Audio. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, 2009.

Kees Van den Doel. Physically based models for liquid sounds. *ACM Transactions on Applied Perception*, 2(4):534–546, October 2005.

Kees Van den Doel and D. K. Pai. Synthesis of Shape Dependent Sounds with Physical Modeling. In *Proceedings of the International Conference on Auditory Displays*, volume 46, 1996.

Kees Van den Doel and Dinesh K. Pai. Jass: a java audio synthesis system for programmers. In *Proceedings of the International Conference on Auditory Display*, Helsinki, Finland, 2001.

Kees Van Den Doel, Paul G. Kry, and Dinesh K. Pai. FoleyAutomatic: physically-based sound effects for interactive simulation and animation. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, Los Angeles, CA, 2001.

Charles Verron and George Drettakis. Procedural audio modeling for particle-based environmental effects. In *Proceedings of the 133rd AES Convention*, San Francisco, CA, October 2012.

Giovanni Vicario. Prolegomena to the perceptual study of sounds. In Davide Rocchesso and Frederico Fontana, editors, *The sounding object*. Mondo estremo, Firenze, 2003.

Jean Vroomen and Beatrice de Gelder. Temporal Ventriloquism: Sound Modulates the Flash-Lag Effect. *Journal of Experimental Psychology: Human Perception and Performance*, 30(3):513–518, 2004.

Marcelo Mortensen Wanderley and P. Depalle. Gestural Control of Sound Synthesis. *Proceedings of the IEEE*, 92(4):632–644, April 2004.

Marcelo Mortensen Wanderley and Nicola Orio. Evaluation of input devices for musical expression: Borrowing tools from hci. *Computer Music Journal*, 26(3):62–76, 2002.

Elisabeth Weis and John Belton, editors. *Film Sound: Theory and Practice*. Columbia University Press, New York, NY, 1985.

David L. Wessel. Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2):45, June 1979.

Ulf Wilhelmsson and Jacob Wallén. A Combined Model for the Structuring of Computer Game Audio. In Mark Grimshaw, editor, *Game Sound Technology and Player Interaction: Concepts and Developments*. IGI Global, September 2010.

Daniel Wilson. *Prehistoric man: researches into the origin of civilisation in the Old and the New world*. Prehistoric man: researches into the origin of civilisation in the Old and the New world. Macmillan, London, UK, 1865.

Luke Windsor. *A Perceptual Approach to the Description and Analysis of Acousmatic Music*. PhD thesis, City University, London, UK, 1995.

Trevor Wishart. *On Sonic Art (Contemporary Music Studies)*. Routledge, July 1996. Published: Hardcover.

Matthew Yee-King and Martin Roth. Synthbot: An unsupervised software synthesizer programmer. In *Proceedings of the International Computer Music Conference*, Belfast, UK, 2008.

John Young. Imagining the Source: The Interplay of Realism and Abstraction in Electroacoustic Music. *Contemporary Music Review*, 15(1-2):73–93, July 1996.

Victor Zappi and Andrew McPherson. Dimensionality and Appropriation in Digital Musical Instrument Design. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 455–460, 2014.

Changxi Zheng and Doug L. James. Toward high-quality modal contact sound. In *ACM Transactions on Graphics (TOG)*, volume 30, page 38. ACM, 2011.

# Appendices

# Appendix A

# Radio Play Script

A radio play was recorded for the second part of the user study presented in Section 4.5. This was based on a script written by David G. Lees and performed by David G. Lees and Stefanie Ritch in December 2013. The script contains nine occurrences of squeaky door sound effects; these are highlighted in bold text.

RADIO PLAY FEATURING THE OCCURRENCE OF NINE SQUEAKY DOORS

Written by David G. Lees (2013)

*A man and a woman arrive at the flat.* ***They open the front door,*** *and walk into the hallway.*

Woman: What a dump!

Man: Well. As I said, it's been years since anyone has been in here.

Woman: ... And did he actually *die* in the house.

Man: I don't know. I don't think so. Let's say no...

Woman: It might be just a touch less creepy if we put some lights on. Is the electricity still connected?

Man: Darling. It's like watching a film with you for the first time. I don't know! Try that little cupboard by the front door. That's where these things usually are.

***Woman opens the cupboard***

Woman: Mmm... They're all flicked up. I suppose that means one of two things... Oh God, I'd hate to get electrocuted. How banal.

Man: You can use a wooden spoon. To flick them the other way.

Woman: Do you have one?

Man: Not on my person, no. But I'm sure there'll be one in the kitchen. Let's have a look.

*They walk slowly through to the kitchen, as if in the dark. The floor is quite creaky.*

Woman: Oh it's lovely! ... Everything just left at is was.

Man: I know, it's too much... God! Look at those glasses! You can't even *buy* sets like that now... Ugh, I don't know why ever it became fashionable to fit kitchens. Aren't they much nicer like this?

Woman: Much... And that FRIDGE! It's a 1940 Kelvinator, you know.

Man: Your encyclopaedic knowledge of refrigerators has always captivated me.

Woman: Don't be sardonic, it's very ugly.

Man: I was trying to be earnest, I know it's much more/

Woman: /Important. I couldn't count the number of times you've said that.

Man: People always mistake my sincerity. It's a curse... See even then!

***Woman opens the fridge***

Woman: Look! It's absolutely FILLED with champagne!

Man: How jolly. Shall we open one?

Woman: Yes, I think we should.

Man: Oh, but it won't be cold.

Woman: People in this country have such an obsession for chilled wines. It isn't right. Do you think Dom Pierre Perignon kept his premier cru at seven to nine degrees?

Man: I suppose not... Although I thought it was ten to fifteen.

Woman: Ten to fifteen? That's practically BATHwater. Anyway, it's pretty well insulated and cool enough in here.

*She pops a bottle*

Woman: Why don't you bring down a couple of those glasses? ... No not the saucers – you're so outré... Ugh no, I've always hated flutes... Just a couple of those tumblers will do. They're far more déclassé...

*Man blows the dust off the glasses and Woman pours two (large) glasses. They cheers.*

Man: Eyes! Eyes!

Woman: Sorry. Cheers.

*They clink glasses again*

Man: Cheers darling.

Woman: Gosh, it's very still in here. So dusty.

Man: I know, it'll be terrible for my asthma.

Woman: You don't have asthma, you smoke too much.

Man: So do you.

Woman: Yes, but I don't blame the after effects on asthma. It's boring. Plus that inhaler you have to carry around and insist on taking after dinner is so squalid.

Man: Mmm *he lights a cigarette.* Open a window then.

***Woman opens a window***

Woman: Oh I just love these old windows. It's so sad when people take them out. I've always thought it would be lovely to have a conservatory made out of these. Isn't the glass beautiful?

Man: Yes, almost sort of rippled. They must be very old... Ah that IS much better. Must be the first fresh in here for nearly 20 years.

Woman: No, it can't be so long!

Man: I told you, it's been like bloody Jarndyce v. Jarndyce.

Woman: How silly it all seems.

Man: Yes, but it's settled now... Sort of... Shall we have a look around?

Woman: Why not?

*They exit the kitchen*

Man: The spoon! *He goes back. He can be heard rustling about in a cutlery drawer. He comes back into the hall.* This'll do.

Woman: That's a spurtle!

Man: Yes, I know – he was always very patriotic in the kitchen.

***He opens the electricity cupboard again and flicks some switches***

Man: Anything?

Woman: Nope, nothing.

Man: Hello darkness my old friend... Well, we've still got some daylight left. Let's look in at the drawing room. I remember there was one lovely picture in there he got from an American chap called Elliot who lived in Paris.

Woman: Which room is it?

Man: The double doors. The out of proportion ones. It was a single, but he had them put in, rather ostentatiously. He didn't even entertain all that much. Well, certainly not the kind of people who would have been impressed by out of proportion double doors.

***She opens the double doors and enters the drawing room***

Woman: What a room! It's so ghostly with all these sheets over the furniture.

Man: Yes – let's take them off. We might as well.

*They take the sheets off the furniture. Probably three. One sofa and two chairs.*

Woman: Oh! These are rather out of line.

Man: Mmm, the Early Dracula Suite we always called it. A man in a bowtie sold it to him, which was very unusual. I think he liked his accent.

Woman: Well, it has a certain...

Man: No, it’s horrible. I hate it and we shall take it into the garden and burn it at the first opportunity.

Woman: Couldn’t we sell it?

Man: Oh, don’t be so middle class about things. No I think a bonfire is just what’s needed. And while we’re at it, I think we’ll burn some of the more hideous paintings too.

Woman: Wonderful, yes – and why don’t we just burn the contents of the library??

Man: You’re exaggerating. They’re only reproductions anyway – the sort you buy on the railings at Hyde Park. I know because one of my horrid uncles came round and pulled the backs off to see if they’d be worth anything. They all have Chinese import stamps on the back. Come on, let’s take them down...

Woman: Ah – and I suppose that’s where the picture he bought from that American chap called Elliot would have been. I always think the marks left by cigarette smoke around where pictures have been are so romantic.

Man: Bugger. Billy must have gotten in here before the lawyers had the place closed up... Give me a hand with this one will you?

*They heave the large painting off the wall*

Man: Shit the *bed!* A wall safe...

Woman: How exciting!

Man: It’ll be 1984. Surely, it has to be...

*He tries the code on the safe. It clicks open*

Man: Oh you predictable old bastard!

***They open the safe***

Woman: Mmm, well not exactly the mines of Solomon...

Man: But what’s this, rolled up at the back?

*He unrolls the scroll*

Man: The picture!

Woman: Darling, that looks like missing East Front of Winchester Cathedral by Turner.

Man: I have *such* a good eye... I think all really refined people have.

Woman: This is a National Treasure.

Man: Like Joan Plowright...

Woman: We should put it back.

Man: No, we should hide it. I always think once a thing has been found that it's sure to be found again if you put it back. Come on…

*They walk through to the bedroom*

Woman: What a beautiful room. Now this is elegant... I don't know what it is...

Man: I think it's the William IV furniture. Very understated – and underrated. Such a perfect bridge from the High Regency style to Victorian. I mean, early Victorian was alright actually – before She went all German Balmorality...

Woman: This linen. Ugh, it's like silk. And the bolsters…*She pats them* If they're not Siberian Goose down, I'm a shrimp... Oh, I could just die...

*She lies down on the bed*

Woman: Ah! And this mattress...

Man: You clever little shrimp indeed... We'll hide the painting under here for a while.

Woman: Under the mattress?!

Man: Yes. It's perfect.

Woman: Darling. Are you kidding? ... That's – well. I mean. That's the most obvious place in the whole world to hide something.

Man: Is it. Don't you think it's sort of - discreet?

Woman: Are you serious? Under the mattress? It's sort of ubiquitous...

Man: Oh. Well I didn't know that. I suppose it must be one of these things that just passed me by. Like, em – honey.

Woman: What do you mean?

Man: Well, my mother never liked honey – so we never had honey. I didn't even know it was a thing until I was seventeen.

Woman: What did you think bees made?

Man: Wax... I tell you. Honey was written out of my narrative. So too it seems has been this universal kink for hiding things under mattresses...

Woman: What a strange fish you are...

Man: So you say probably it's not a good idea?

Woman: No dear. Let's just keep it with us for the time being.

***A door unexpectedly opens, and a stranger enters the room***

Man: Good Christ! Where did you come from?

Kit: There's a connecting passage between this apartment and mine. A charming little baize lined affair, Eddy and I used it to share staff you see. They got so expensive after the war.

Man: Who are you?

Kit: Kit Douglas, Viscount Drumlanrig. How do you do?

Man: How do you do?

Woman: How do you do?

Kit: How do you do? ... May I be so bold as to ask what you're doing here?

Man: Well I own the place now, actually.

Kit: How interesting, it'll be nice to have a neighbour again. Took rather a while coming through, didn't it?

Man: Yes – all rather messy I'm afraid. People who you've always thought so perfectly nice can get so wretched over money.

Kit: Indeed... I wonder, do you have some paperwork to prove who you are? I... Well, if you are the rightful heir, there are some things I should return to you.

Man: I.. Um, yes... The letters from the lawyers.

*Man shows him the lawyers letters*

Kit: Well well well. Edith's son?

Man: Yes, Eddy was my mother's brother.

Kit: How is dear Edith?

Man: She died, six years ago. She had a stroke.

Kit: I'm so sorry to hear that. It must have been very difficult for you. Edith and I were very close once...

Woman: Shall I bring through some champagne?

Kit: From where?

Woman: The refrigerator is full of it.

Kit: Hah! If only I'd known... Well yes, let's have a little drink.

*Woman exits*

Kit: You have very similar features to Eddy. Sympatico, I'd say.

Man: That’s very kind of you. I didn’t know him well, I was too young. I suppose you must have done.

Kit: Oh yes, very well. I miss him dreadfully.

*Woman returns with champagne and another glass. She fills all three.*

Kit: To Eddy.

Man & Woman: To Eddie

*A pause*

Kit: Now, those things I was telling you about.

Man: Oh yes, of course.

Kit: Could you give me a hand with this mattress?

Woman: Hah!

Kit: What’s so funny?

Woman: I was saying just before you came in, no one really hides things under mattresses.

Kit: Really? ... Oh I’ve always considered it rather discreet.

*Some heaves while they move the mattress*

Woman: Good grief!

Kit: Yes, quite a little trove. Nothing valuable I’m afraid.

Man: But some lovely things. Who’s that, in the kilt?

Kit: That’s your great grandfather, Lachlan Campbell. The umbrella stand behind him is around here somewhere. I hid it just after Eddy died, when Billy was snooping around. Would’ve been terribly sad if he’d gotten his hands on it... I’d love to talk you through the whole lot. But why don’t we have lunch first?

Woman: I’m absolutely starving, that sounds perfect.

Kit: Fine, I’ll take you round to my club. I think you’ll like it. The food is tolerable, chop house sort of affair, but the environment ‘s the thing. And one can still smoke inside.

Woman: Wondair!

Man: Let us go then...

Kit: I haven’t been out through the front door here for over twenty years.

***They exit, opening the large iron storm grate***

Kit: always forget how pigging heavy that thing is…

END

# Appendix B

# Foley Questionnaire and Supplementary Diagrams of Results

## B.1 Questionnaire

An online questionnaire[1] was conducted with Foley artists, as discussed in Section 5.2. The full questionnaire is presented on the following pages.

[1] hosted on http://www.soscisurvey.de

Queen Mary
University of London

**1. Age**

- ○ 15-20
- ○ 20-25
- ○ 25-30
- ○ 30-35
- ○ 35-40
- ○ 40-45
- ○ 45-50
- ○ 50-55
- ○ 55-60
- ○ 60-65
- ○ 65-70
- ○ 70+

---

- ○ Prefer not to say

**2. Gender**

- ○ Male
- ○ Female

---

- ○ Prefer not to say

**3. Continent**

- ○ Africa
- ○ Asia
- ○ Australia
- ○ Europe
- ○ North America
- ○ South America

**4. How many years of professional experience do you have working as a Foley artist or sound designer?**

- ○ Less than 1 year
- ○ 2-5 years
- ○ 5-10 years
- ○ More than 10 years

- ○ 1-3 projects
- ○ 4-10 projects
- ○ 10-20 projects
- ○ More than 20 projects

**6. What types of projects have you worked on professionally?**
You can select multiple types

- ☐ Short film
- ☐ Feature film
- ☐ Animation
- ☐ Games
- ☐ Theatre
- ☐ Other

**7. On average, how much of your time on projects is spent performing Foley to on-screen (or on-stage) actions?**

Percentage of time spent performing Foley

**8. Have you done any Foley for games or interactive experiences?**

- ○ Yes
- ○ No

**9. Do you play a musical instrument?**

- ☐ Bowed string instrument
- ☐ Non-bowed string instrument
- ☐ Percussion
- ☐ Brass
- ☐ Woodwind
- ☐ Keyboard
- ☐ Electronic
- ☐ Other

**10. How many years of experience do you have playing your main musical instrument?**

- ○ Less than 1 year
- ○ 1-2 years

- ○ 5-10 years
- ○ More than 10 years

---

- ○ I don’t play any musical instruments

**11. If you participated in the study at QMUL, please enter your first name:**

Back Next

---

Pause the interview 17% completed

Queen Mary
University of London

This section is about your experience in handling sounding objects (referred to here as 'props') while synchronising sound. Props include all objects that you use to perform sounds in sync to the moving image.

**12. Approximately, what percentage of the props you use for professional work are...**

Estimate a percentage for each of the following cases. Total percentages don't need to add up to 100%.

encountered for the first time

used regularly across many different projects

**13. Respond to the following statements.**

| | strongly disagree | disagree | neither agree nor disagree | agree | strongly agree | don't know |
|---|---|---|---|---|---|---|
| I generally find it easy to recreate the same take using the same prop. | ○ | ○ | ○ | ○ | ○ | ○ |
| I spend a lot of time exploring the different *timbral characteristics* of a prop. | ○ | ○ | ○ | ○ | ○ | ○ |
| I spend a lot of time exploring different *ways of handling and performing* a prop. | ○ | ○ | ○ | ○ | ○ | ○ |
| I often get frustrated with some props because they are unable to produce the sound I want them to. | ○ | ○ | ○ | ○ | ○ | ○ |
| When handling a prop I have a clear idea of the sound I want to produce with it. | ○ | ○ | ○ | ○ | ○ | ○ |
| I often have trouble recreating a certain sound or sequence from a previous take. | ○ | ○ | ○ | ○ | ○ | ○ |
| I am often surprised by the sounds that a prop is capable of producing. | ○ | ○ | ○ | ○ | ○ | ○ |
| I try recording a take before exploring the prop's timbre. | ○ | ○ | ○ | ○ | ○ | ○ |
| I often get frustrated with some props because they are difficult to perform with. | ○ | ○ | ○ | ○ | ○ | ○ |

**15. Respond to the following statements.**

| | strongly disagree | disagree | neither agree nor disagree | agree | strongly agree | don't know |
|---|---|---|---|---|---|---|
| I focus more on the on-screen (or on-stage) action than on the timbral nuances of the sound | ○ | ○ | ○ | ○ | ○ | ○ |
| Handling props is much like playing a musical instrument | ○ | ○ | ○ | ○ | ○ | ○ |
| Generally, my actions and movements when performing Foley are very similar to the actions and movements depicted on screen | ○ | ○ | ○ | ○ | ○ | ○ |
| When performing with a prop I pay a lot of attention to timbral nuances of the sound | ○ | ○ | ○ | ○ | ○ | ○ |
| My movements or actions while handling props often bear little resemblance to the on-screen (or on-stage) action being dubbed | ○ | ○ | ○ | ○ | ○ | ○ |

| | strongly disagree | disagree | neither agree nor disagree | agree | strongly agree | don't know |
|---|---|---|---|---|---|---|
| In Foley, behaviour is more important than sound quality | ○ | ○ | ○ | ○ | ○ | ○ |
| I would rather use a high-quality sound that doesn't match the on-screen (or on-stage) action than have a low-quality sound over which I have full control | ○ | ○ | ○ | ○ | ○ | ○ |
| The human-performed aspect of Foley causes the dubbed soundtrack to be less than ideal | ○ | ○ | ○ | ○ | ○ | ○ |

**16. Do you think Foley can be expressive (i.e. aesthetically meaningful)? If so, can you comment on what you think makes a Foley track expressive rather than a practical necessity?**

**17. Do you have any further comments on this section?**

Queen Mary
University of London

This section is about the relationship between sound and image in Foley synchronisation.

**Important: this section refers exclusively to sounds that require being performed while viewing the on-screen (or on-stage) action. (i.e. this excludes one-off sounds like gunshots that don't require viewing the image for performing the alignment of action to sound)**

| | strongly disagree | disagree | neither agree nor disagree | agree | strongly agree | don't know |
|---|---|---|---|---|---|---|
| In physical terms, the sounds I produce are usually exaggerated versions of what is depicted on screen (or on stage) | ○ | ○ | ○ | ○ | ○ | ○ |
| Even if the sounds I make are physical exaggerations of the on-screen (or on-stage) action, I try to maintain consistency for similar actions | ○ | ○ | ○ | ○ | ○ | ○ |
| I often sacrifice physical consistency in order to achieve more accurate temporal synchronisation | ○ | ○ | ○ | ○ | ○ | ○ |

| | strongly disagree | disagree | neither agree nor disagree | agree | strongly agree | don't know |
|---|---|---|---|---|---|---|
| I rely on post-editing to get the timing of my performance right | ○ | ○ | ○ | ○ | ○ | ○ |
| For aesthetic reasons I often intentionally perform sounds too late or too early compared to their visual counterpart | ○ | ○ | ○ | ○ | ○ | ○ |
| Temporally speaking, the Foley track does not need to be perfect | ○ | ○ | ○ | ○ | ○ | ○ |
| I often sacrifice temporal accuracy in order to improve other aspects of the sound | ○ | ○ | ○ | ○ | ○ | ○ |
| Precise timing only matters at specific moments in the shot | ○ | ○ | ○ | ○ | ○ | ○ |

**18. Can you think of any specific cases where you have used temporal or physical discontinuities between the sound and the image as an aesthetic choice?**

| | strongly disagree | disagree | neither agree | agree | strongly agree | don't know |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| I am normally given a project brief including narrative and/or aesthetic details before starting a project | ○ | ○ | ○ | ○ | ○ | ○ |
| Narrative and/or aesthetic details have a large impact on the way I perform Foley | ○ | ○ | ○ | ○ | ○ | ○ |

**19. On average, how many takes do you normally do for a shot or scene before being satisfied with the result?**

- ○ 1-2
- ○ 2-5
- ○ 5-10
- ○ 10-20
- ○ More than 20

**20. Please rank your most common reasons for rejecting a take.**

| | | |
|---|---|---|
| **Timing** | **Timbral detail** | **1** |
| **Ergonomics (e.g. prop slipping out of hands)** | **Technical fault** | **2** |
| **Other** | | **3** |
| | | **4** |
| | | **5** |

**21. Do you have any further comments on this section?**

Back Next

Pause the interview 50% completed

Queen Mary University of London

**22. How many games or interactive media projects have you done Foley work for?**

- ○ 1-2
- ○ 3-5
- ○ More than 5

**23. What material are you given when doing Foley work for games or other interactive media?**

- ☐ Example gameplay sequences
- ☐ Cut-scenes (non-gameplay sequences)
- ☐ Concept art
- ☐ Other

**24. In your opinion, how successfully has your Foley been integrated into games or interactive experiences?**

- ○ Very well
- ○ Well
- ○ Ok
- ○ Not very well
- ○ Terribly

- ○ I didn't get to experience the end result
- ○ My Foley wasn't integrated into interactive parts of the game or experience

**25. Do you have any further comments about your experiences with games (e.g. differences to traditional Foley work)?**

Back Next

Pause the interview 67% completed

Queen Mary University of London

**26. This question is about your experience using digital interfaces. Note that this excludes recording equipment.**

| | strongly disagree | disagree | neither agree nor disagree | agree | strongly agree | don't know |
|---|---|---|---|---|---|---|
| I have experience using digital interfaces for post-production (e.g. MIDI controllers, joysticks, sensors) | ○ | ○ | ○ | ○ | ○ | ○ |
| I regularly use digital interfaces to perform Foley | ○ | ○ | ○ | ○ | ○ | ○ |

| | strongly disagree | disagree | neither agree nor disagree | agree | strongly agree | don't know |
|---|---|---|---|---|---|---|
| There won't be any need for Foley artists in the future because their work will be automated by computers and algorithms | ○ | ○ | ○ | ○ | ○ | ○ |
| Future Technology will be enabling for Foley artists | ○ | ○ | ○ | ○ | ○ | ○ |
| I am skeptical about using digital interfaces to do Foley | ○ | ○ | ○ | ○ | ○ | ○ |
| I am skeptical about real objects being replaced by synthesisers (e.g. physical models) in the future | ○ | ○ | ○ | ○ | ○ | ○ |

**27. Would you like to leave any further comments about the implications of future technology on Foley?**

Back Next

Pause the interview

83% completed

## B.2 Supplementary Diagrams of Results

Supplementary diagrams of results from the online questionnaire presented in Section B.1 of this appendix are presented here. These are referred to and discussed in Section 5.2.

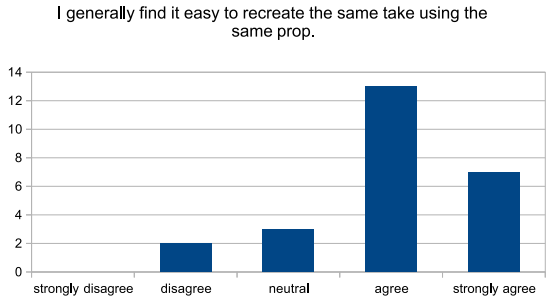


Figure B.1: Likert-scale responses for Question 13(a)

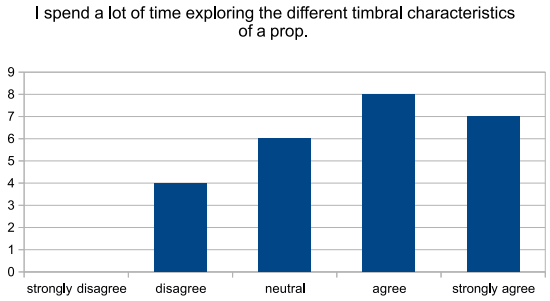


Figure B.2: Likert-scale responses for Question 13(b).

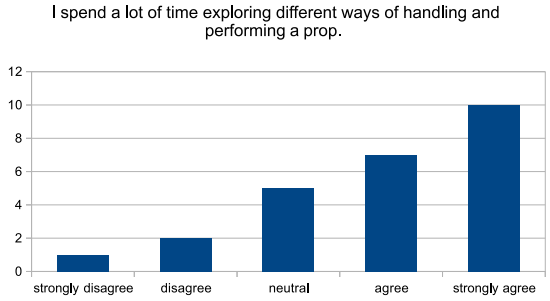


Figure B.3: Likert-scale responses for Question 13(c)

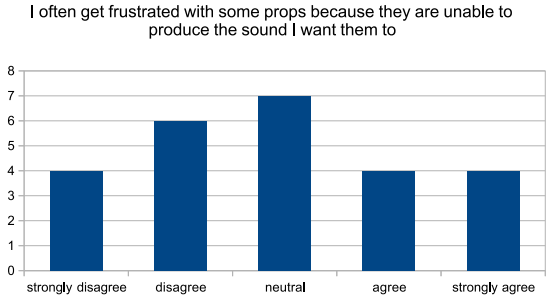


Figure B.4: Likert-scale responses for Question 13(d)

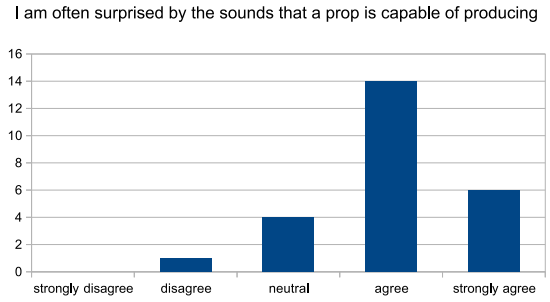


Figure B.5: Likert-scale responses for Question 13(g)

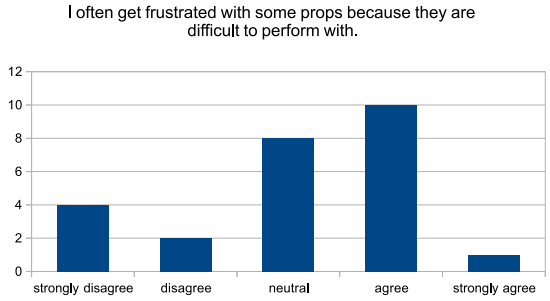


Figure B.6: Likert-scale responses for Question 13(i)

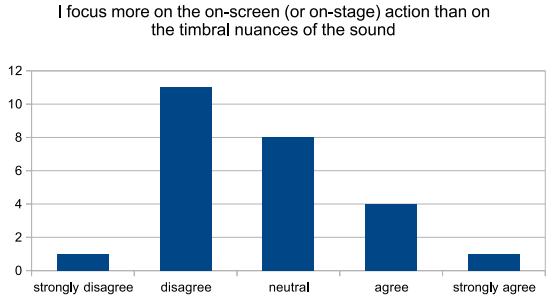


Figure B.7: Likert-scale responses for Question 15(a))

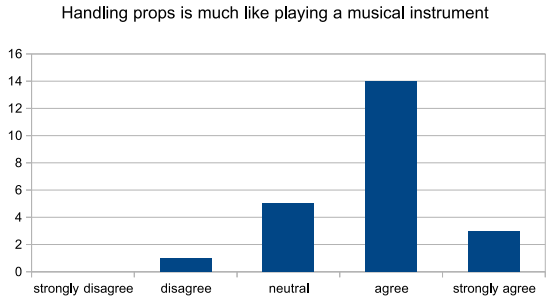


Figure B.8: Likert-scale responses for Question 15(b)

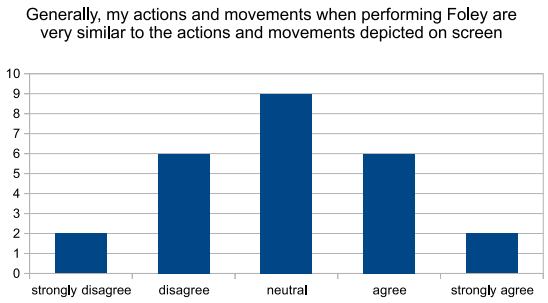


Figure B.9: Likert-scale responses for Question 15(c)

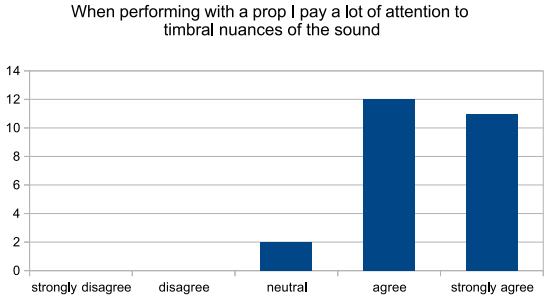


Figure B.10: Likert-scale responses for Question 15(d)

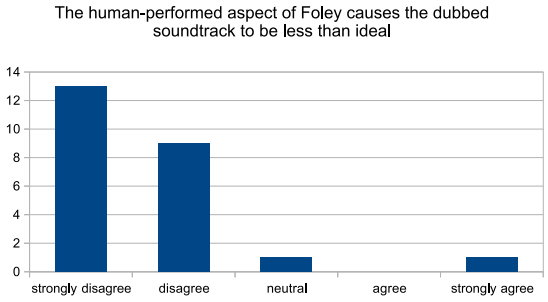


Figure B.11: Likert-scale responses for Question 15(h)

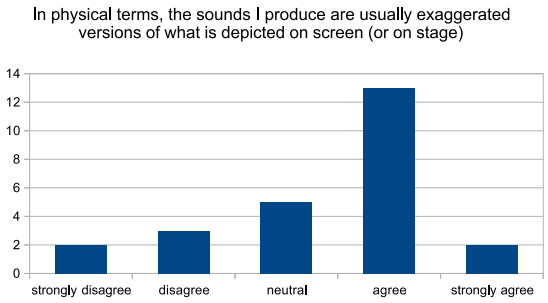


Figure B.12: Likert-scale responses for Question 17(a)

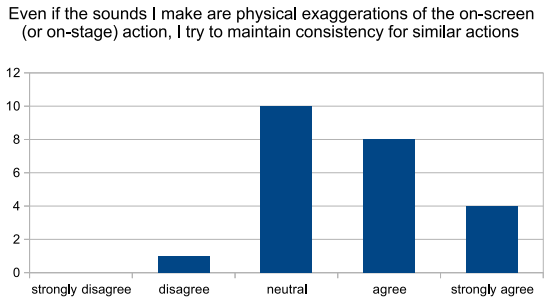


Figure B.13: Likert-scale responses for Question 17(b)

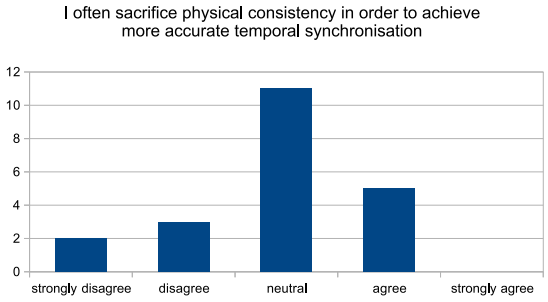


Figure B.14: Likert-scale responses for Question 17(c)

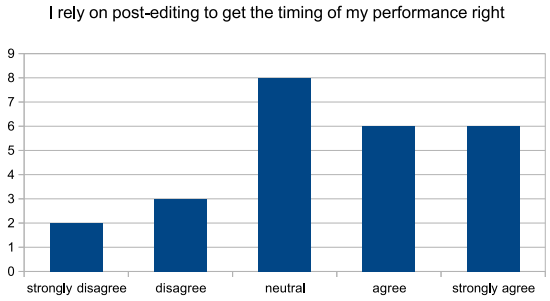


Figure B.15: Likert-scale responses for Question 17(d)

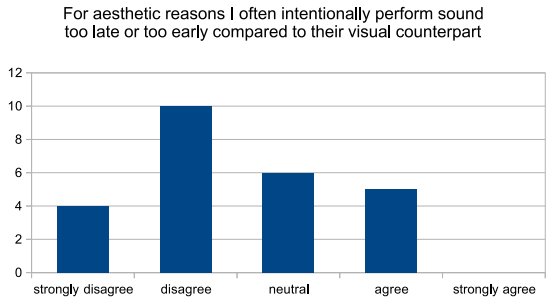


Figure B.16: Likert-scale responses for Question 17(e)

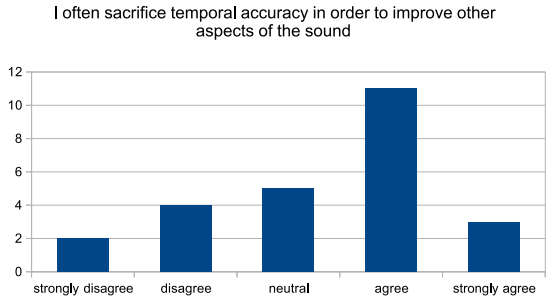


Figure B.17: Likert-scale responses for Question 17(g)

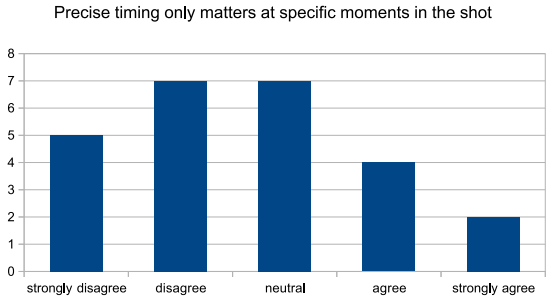


Figure B.18: Likert-scale responses for Question 17(h)

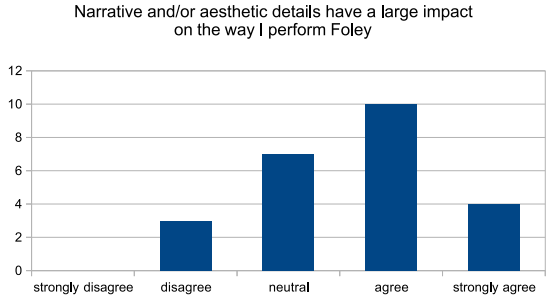


Figure B.19: Likert-scale responses for Question 18(b)

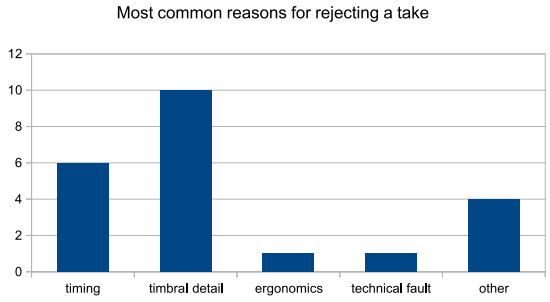


Figure B.20: Rankings for Question 20

# Appendix C

# Animation Screenplay for Synchronisation Study

The audio-visual synchronisation study described in Chapters 5 and 6 was based on a screenplay, presented on the following pages. Participants of the study were given a copy of this screenplay before watching the corresponding animation.

SCREENPLAY
Nailex's Big Break (W.T.)

## SCENE 1 - GYMNASIUM

Nailex is in a gymnasium in front of a large ring standing on the floor in front of him. He is practising for an important acrobatics show in a couple of days time.

He looks up at the ring, down at the marking on the floor, and then starts walking backwards, carefully measuring the runway up to the ring. He proceeds to practise the jumping through the ring a few times, more or less successfully, though sometimes overstepping the markings or misjudging the landings.

## SCENE 2 - STAGE

Nailex is at the entrance to a large stage. He confidently struts onto the stage towards the centre.

[SFX: APPLAUSE]

As the camera follows Nailex to the stage, a large ring can be seen, on fire, suspended in the air over a large gap in the stage.

[SFX: APPLAUSE STOP SLOWLY]

Nailex doesn't notice this straight away and only looks up as he approaches the edge of the stage. He looks up at the ring, down the gap, gets a fright and turns around frantically, running back to the entrance of the stage. He turns around to face the stage again.

## SCENE 3 - STAGE (DREAM SEQUENCE)

Nailex can be seen successfully jumping across the ledges through the ring while performing somersaults and graceful landings. As he jumps through the ring, stage lights flash along with loud applause from the audience.

MONTAGE:
4 scenes, jumping right-to-left or left-to-right.
A, B, C, D

[SFX: APPLAUSE]
[SFX: LIGHTS SWITCHING ON]

## SCENE 4 - STAGE

Fade back to the end of scene 2. Nailex is facing the stage. He suppresses his fears and strides towards the ledge. As he approaches the ledge he prepares to jump, raising his arms in the air, but fails to take the jumping leap on time and falls across the ledge. His limbs rotate around their joints, occasionally hitting the walls (sides of the stage) as he falls down the dark pit. Fade out.

## SCENE 5 - WORKSHOP

Fade into a workshop bench with several human-sized tools and a desk lamp. After a couple of seconds Nailex can be seen falling from above into the scene and onto the workbench. He lands awkwardly, one of his limbs rocking helplessly back and forth. He slowly gets up, limps a bit and then shuffles towards the left of the scene, exits the scene. Camera is stationary. Fade out.

# Appendix D

# *FoleyDesigner* Prototype

This appendix outlines the *FoleyDesigner* project, funded by Queen Mary Innovation[1] and developed over a five-month period between May and September 2015. The aim of this project was to develop a combined hardware and software prototype that incorporates human performance into the development of computational sound models for games. The project arose directly out of work carried out for this thesis, particularly the use of hardware sensors to control timbre-led models in Chapters 3 and 4. Some integration techniques borrowed from the field of graphical animation were also incorporated here, including keyframe-based behaviour interpolation discussed in Chapter 7.

## D.1 Overview of System

The proposed workflow involves processing real-time sensor inputs using so-called *control layers* to drive parameters of a computational sound model. Control layers can read several streams of data from physical sensors (e.g. accelerometers, capacitive touch sensors, potentiometers) and process them according to a user-defined transformation. Examples of transformations include smoothing, differentiation and more complex ones such as friction simulation. Control layers can be re-used for different sensor configurations and projects, and can easily be stored as templates or examples for later implementation. Control layers can also generate their own data without the use of a physical input. Outputs from these control layers can be routed to the parameters of a sound model in real-time without interrupting the audio output, allowing for quick exploration and experimentation.

The sound model can be chosen from a pre-existing library, be extended by the user or programmed by the user. The sound model need not adhere to physical

[1] http://www.qminnovation.co.uk

principles and can instead have abstract parameter input descriptions such as *pitch*, *roughness*, *brightness*, and so forth. The user can perform behaviours (i.e. parameter trajectories) using the physical sensors. Performances can be recorded at any time. Once recorded, all corresponding sensor data is stored into a library of animations, which can be processed further by the user. The sensor data is converted from a regularly sampled stream into a series of keyframes (also known as 'breakpoints') in order to maintain a small file size and enable more natural blending techniques between animations (described below). Processing of recorded sensor data involves smoothing the data and setting start and termination points. In addition, the user can choose to repeat the keyframe reduction process by specifying the amount of desired keyframes per second. Animations can be auditioned at any time during this process.

After accumulating the desired amount of animations, the sound model and animation library can be exported into a standalone plugin. An exporting tool allows the user to define a set of meta-parameters which are used to automatically playback and blend animations. Multiple animations can be blended in real-time according to animation weightings specified by the user. Keyframe data is warped on both temporal and vertical axes, allowing for natural transitions from one configuration to the next. The chosen *meta-parameters* are exposed in the standalone plug-in and can easily be driven from a game engine or any other interactive environment. See Figure D.1 below for a visual representation of the system.

## D.2 Interaction with Bela Embedded Audio Platform

Bela[2] is an embedded audio platform based on the Beaglebone Black and developed at the Centre for Digital Music, Queen Mary University of London (McPherson and Zappi, 2015; Moro et al., 2016). The FoleyDesigner software communicates with the Bela board via scripts and UDP network messages. The user is able to address each analog input and output, as well as stereo audio input and output on the board. This is done using Puredata patches which are converted into Vanilla C code using the Heavy Cloud Compiler (described below). The generated code is used in conjunction with specially developed Bela code which enables the user to easily address inputs and outputs of the board using conventional Puredata objects that wouldn't otherwise be used for this purpose. The software can automatically generate Puredata patches, upload patches to Enzien Audio's compiler servers, send the generated code to the BeagleBone Black (plus Bela cape), interface with Bela (e.g. to let generated code be compiled on the board), run the generated executable on the board and monitor its

[2] http://bela.io

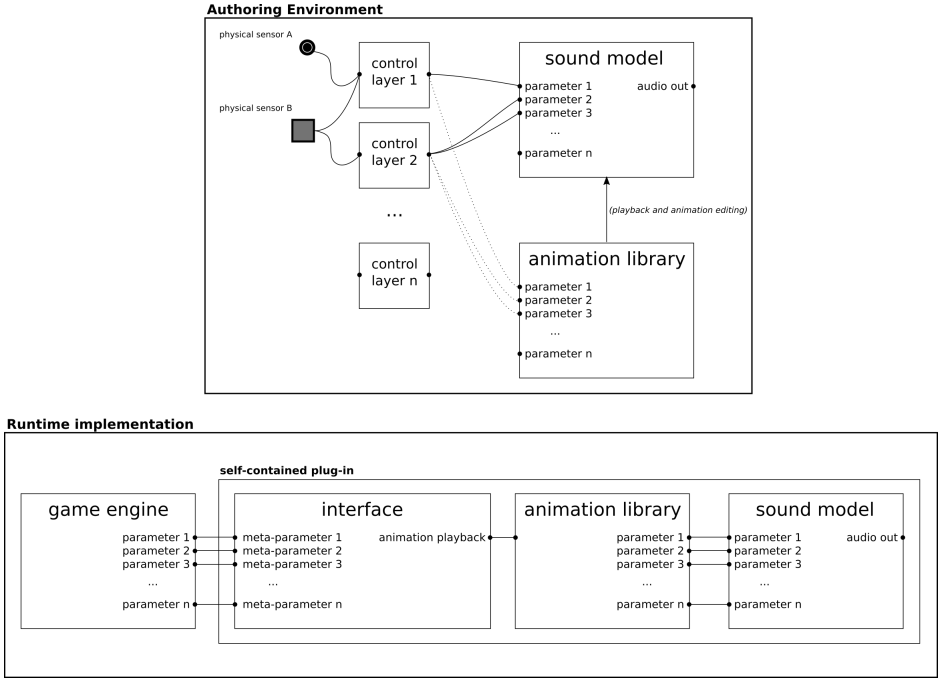


Figure D.1: Block diagram illustrating the authoring and run-time components of the FoleyDesigner prototype.

output without the user having to leave the software environment. For each sensor control layer generated by the user, the output is monitored by the Bela program and sent to the FoleyDesigner software running on the host machine, which in turn visualises the data in the style of an oscilloscope. The user is able to interact with the visualisation with the mouse in order to zoom in and scroll. Audio output RMS values are also sent back to the host machine in order for the user to monitor the output level. The hardware prototype consists of six sensors attached to the analog inputs of the Bela platform: an accelerometer, a force-sensitive resistor, two piezo discs and two linear faders (see Figure D.2).

## D.3 Interaction with Puredata and Enzien Audio's Heavy Cloud Compiler

Enzien Audio's Heavy Cloud Compiler[3] is a service that allows a user to upload Puredata patches and receive optimised C code (amongst other outputs) in return. The sound model and control layers specified by the user are expressed as Puredata patches with the aid of specially developed Puredata abstractions that are designed to present

[3]https://enzienaudio.com

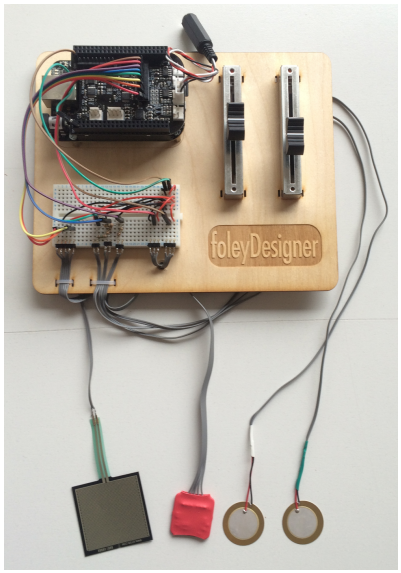


Figure D.2: One of the assembled hardware kits, based on the Bela audio and sensing platform.

the user with a more readable layout. After the user has configured the sound model and control layers (either from within the software or using Puredata), the software can process the corresponding Puredata patches and generate new ones, removing user- friendly abstractions and replacing them with all the necessary code to produce the expected interaction with the Bela platform's features. The generated patch is then sent to Enzien Audio's servers via a Python script (developed and supplied by Enzien Audio) and C code compatible with the Bela platform is received in return. This code is then copied over to the board, compiled on the board and launched. The whole process takes approximately between 5 and 30 seconds, depending on the complexity of the synthesiser and control layers designed by the user.

## D.4 Control Layers and Synthesis

The content generated by the user consists of control layers and a synthesiser. A control layer consists of a Puredata patch that takes sensor readings and processes them to produce a single continuous audio output (44.1kHz sampling rate and 16bit resolution). The synthesiser is also a Puredata patch which produces sound based on the values of public parameters (exposed to the FoleyDesigner interface using specially developed Puredata abstractions). Once the FoleyDesigner session has been compiled on the board, the user can route the outputs of each control layer to the public parameter inputs of the synthesiser using the provided user interface (see Figure D.3).

A public parameter can also be given a constant value rather than be controlled by the output of a control layer. This is all done in real-time without needing to re-compile the project.

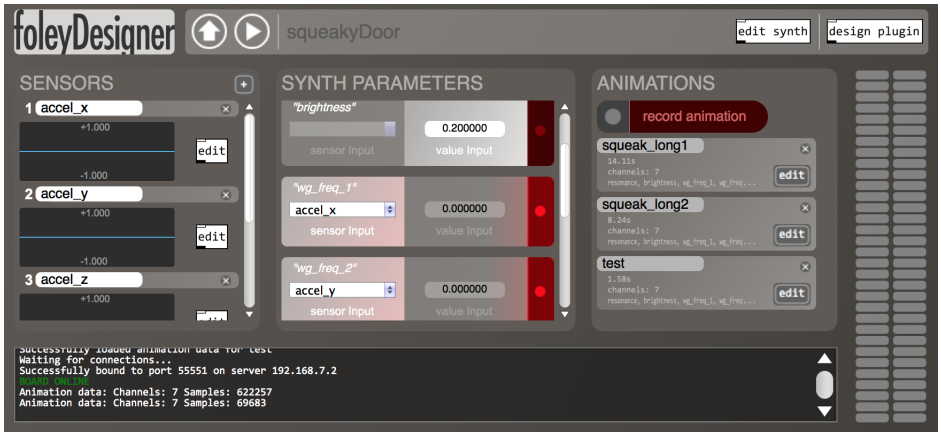


Figure D.3: *FoleyDesigner* main user interface.

## D.5 Recording of Gestures and Transformation into Keyframed Animations

Each of the public parameters can be enabled for recording. Upon toggling a global 'record' button, all inputs to synthesiser parameters (including static values) are recorded on the board (as raw binary data at audio sampling rate and resolution). When finished recording, the data is sent back to the host machine. The FoleyDesigner software then processes this data and represents it as a so-called *animation* on the right hand side of the interface. The user can choose to edit this animation using the *animation editor* (see Figure D.4). The animation editor contains two viewing methods: a *track view* and a *master view*. In the master view the user is presented with an overview of all the tracks and their corresponding data and can select the start and end-points of the animation. The animation can also be made *seamless*, by cross-fading a specified length of time at the beginning and end of the time selection (making it possible to loop the recorded data without hearing any transients at the looping point). In the track view each individual track of recorded data can be edited. The data can be smoothed (by applying a moving-average filter), and reduced to keyframes. Keyframes are two-dimensional vectors that represent a value and a time. They are commonly used in animation software and are often referred to as *breakpoint envelopes* in audio workstations. The user can set the desired density of

keyframes (in units of keyframes per second). The animation editor also provides information to the user such as the duration and amount of keyframes in the animation.

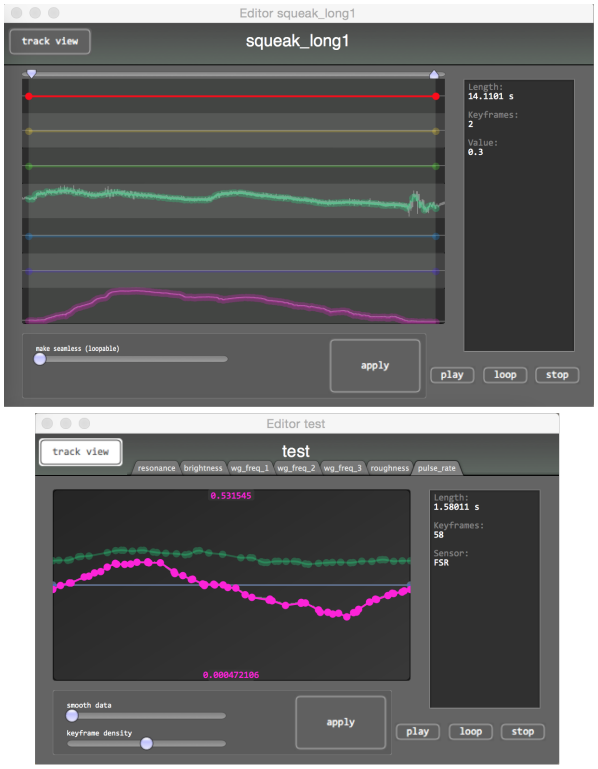


Figure D.4: The animation editor window: *master view* (above) and *track view* (below). Recorded animations are displayed on a per-parameter basis as keyframe data. The editor lets the user process the recorded data by applying smoothing and reducing the amount of keyframes.

## D.6 Exporting of Audio Plugin and Keyframe Data

Keyframe data can easily be exported to XML data, allowing it to be used in common animation software, audio middleware and game engines. The synthesiser itself can be exported to an audio plugin (VST, AU, Wwise, Unity) using Enzien Audio's cloud compiler. The software is also capable of generating a further Puredata patch that contains the synthesiser as well as the animation data. The Puredata patch is designed to be opened on the host machine (rather than compiled onto the Bela platform, though this is also possible). Within this patch the user is able to playback and blend recorded animations. Animation times and values are stored in separate tables making it possible to interpolate multiple parameter trajectories in both dimensions of

time and value. Multiple animations can be interpolated at any given time. The generated patch allows the user to easily generate new parameters (or *meta-parameters*) to control the weighting of animations and their playback. After creating a meta-parameter the user can store the state of the patch (i.e. the individual weightings of the animations) by clicking a button corresponding to the lower or upper value of a meta-parameter. Animations can be played back in three different ways: *trigger*, *loop* or *scrub*. In *trigger* mode, the patch contains a button-style parameter to play back the (blended) animation from beginning to end. In *loop* mode, the patch contains a toggle-style parameter to toggle playback, and a continuous parameter controlling the speed of the playback. In *scrub* mode, a continuous parameter controls the time position of the blended animation.

## D.7 Case Study

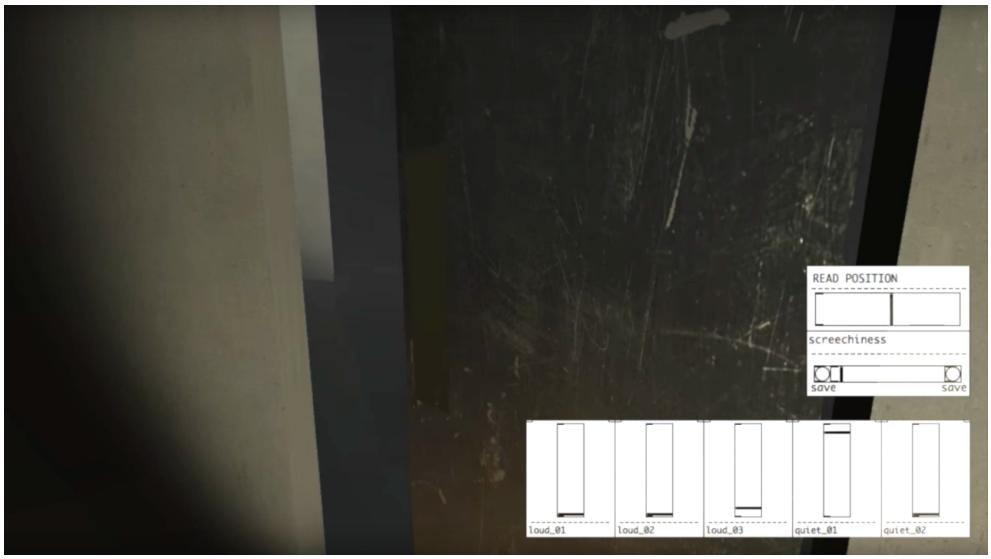


Figure D.5: An example implementation of a sound model authored in *FoleyDesigner*. Angular velocity and angle of the door are mapped to meta-parameters (*screechiness* and *read position*) controlling performed animations.

Figure D.5 shows a screenshot of a complete implementation of a sound model authored in FoleyDesigner. An accelerometer and a force-sensitive resistor were used to perform a set of behaviours on a friction model. Weighting presets for these animations were mapped to a meta-parameter labelled *screechiness*. A further meta-parameter controlled the *read position* of each animation (normalised to the range [0-1]).

An interactive door (implemented in the Unity game engine) was used to control this sound model. The angle of the door was mapped to the read-position of the animations, and its angular velocity was mapped to the *screechiness* meta-parameter. This resulted in an interaction where fast movements would result in a louder sound than attempts to open the door slowly.

## D.8 Presentations

The *FoleyDesigner* prototype was presented at the *Game Developer Conference*[4] in March 2016 and at the *AES Audio for Games Conference*[5] in February 2016.

## D.9 Video

A demonstration video of the prototype can be found in the on-line supplementary audio-visual materials (see Appendix E.4).

[4]http://www.gdconf.com
[5]http://www.audioforgames.co.uk/2016

# Appendix E

# Supplementary AV Materials

Supplementary audio-visual materials referenced in this document are available online under the following URL:

`http://www.eecs.qmul.ac.uk/~andrewm/heinrichs_thesis/`

## E.1 Reference Sounds and Model Outputs in Development of Creaking Door Model

Original reference sounds and versions emulated through the creaking door model at each stage of its design process, described in Section 4.3.2.

## E.2 Control Strategies for Performing the Creaking Door Model

A video demonstrating each of the control layers for performing the creaking door model, as described in Section 4.4.3.

## E.3 Physical Reference and Performed Soundtracks from Synchronisation Study

Videos of the animation, *Nailex's Big Break*, with each of the participants' soundtracks and the physical reference soundtrack, discussed in Chapter 6. The videos contain overlays of frame numbers relative to each scene.

## E.4 FoleyDesigner Overview

A video demonstrating the *FoleyDesigner* prototype presented in Appendix D.